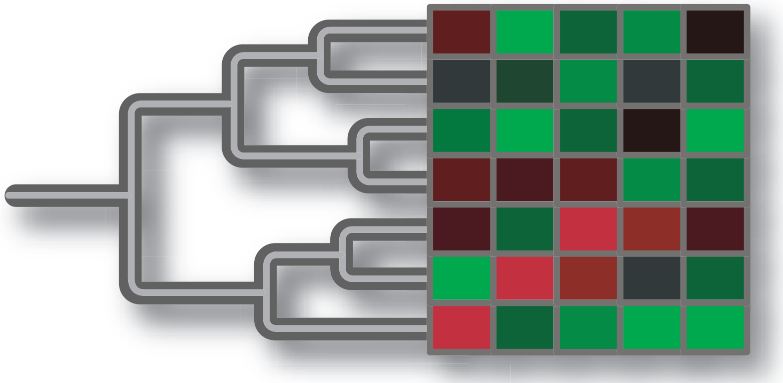
GE Healthcare

DeCyder Extended Data Analysis module Version 1.0

Module for DeCyder 2D version 6.5

User Manual



Contents

9 Calculation and Results - Pattern Analysis

10 Calculation and Results - Discriminant Analysis

11 Interpretation

12 Creating and managing sets

13 Exporting data from EDA

14 Tutorials Introduction

15 Tutorial I - Identify spots for picking and import MS data

Appendix E Statistics and algorithms - Pattern Analysis

Appendix F Statistics and algorithms - Discriminant Analysis

Appendix G DiscoveryHub

1 Introduction

1.1 Introduction

DeCyder™ Extended Data Analysis Software (denoted EDA in the manual) is a high-performance proteomics informatics software for analysis of large and combined data sets. EDA was developed specifically for the 2D DIGE methodology and therefore all the advantages of this approach are utilized in the software.

EDA is an add-on module for the DeCyder 2D Software. It is used for multivariate analysis of protein expression data derived from the BVA module or the Batch Processor. EDA can handle up to 1000 spot maps. The raw data (gel images) are linked to EDA and can be opened for display via the BVA module.

In addition to the univariate analyses (Student's T-test, One-way ANOVA and Two-way ANOVA) that can be performed in the BVA module, it is also possible to perform the following analyses in EDA:

- **Principal Component Analysis**
  Produces an overview of the data. Can be used to find outliers in the data.

- **Pattern analysis**
  Finds patterns in expression data (e.g. proteins and spot maps with similar expression profiles).

- **Discriminant analysis**
  Finds proteins that discriminate between different samples (to find biological markers), creates classifiers and assigns samples to known classes depending on expression profiles (e.g. tumor typing).

- **Interpretation**
  Finds the biological context of proteins by integrating biological information and context from in-house or public databases. It can be used to determine in what pathways and processes a protein is involved, the function of the protein etc.

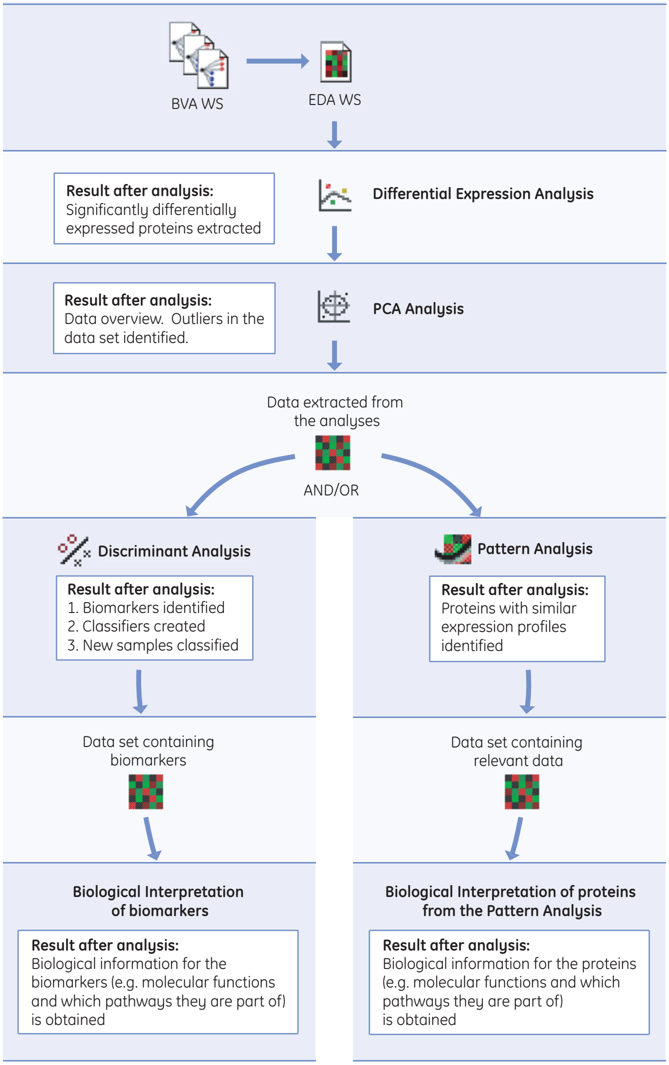

BVA WS
EDA WS
Result after analysis:
Significantly differentially expressed proteins extracted
Differential Expression Analysis
Result after analysis:
Data overview. Outliers in the data set identified.
PCA Analysis
Data extracted from the analyses
AND/OR
Discriminant Analysis
Result after analysis:
1. Biomarkers identified
2. Classifiers created
3. New samples classified
Pattern Analysis
Result after analysis:
Proteins with similar expression profiles identified
Data set containing biomarkers
Data set containing relevant data
Biological Interpretation of biomarkers
Result after analysis:
Biological information for the biomarkers (e.g. molecular functions and which pathways they are part of) is obtained
Biological Interpretation of proteins from the Pattern Analysis
Result after analysis:
Biological information for the proteins (e.g. molecular functions and which pathways they are part of) is obtained

Fig 1-1. Simplified overview of calculations in EDA.

1.2 The DeCyder EDA User Manual

This user manual is broadly divided into 3 main parts: the reference manual (Chapters 1-13), the tutorials (Chapters 14-16) and the appendices.

It is recommended that new users first work through the tutorials, in order to gain a rapid understanding of the software's capabilities. The tutorials are step-by-step guides that take the user through the main applications of the software by employing real data. The tutorials will be provided on a DVD and must be imported into the database (see Chapter 14 for more information). The tutorials are designed to be worked through without prior knowledge of the reference component of the manual but with knowledge of the DIGE, DIA and BVA concepts.

The reference manual provides a more detailed technical account encompassing important aspects of the built-in functionality of EDA, which can be used as a source of further information for experienced users.

Specific details of for example normalization and the statistical analyses can be found in the appendices.

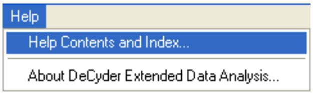
For further help with details see the EDA online help. It can be accessed from the software in several ways. See section 1.3, Getting help for more information on how to access the help.

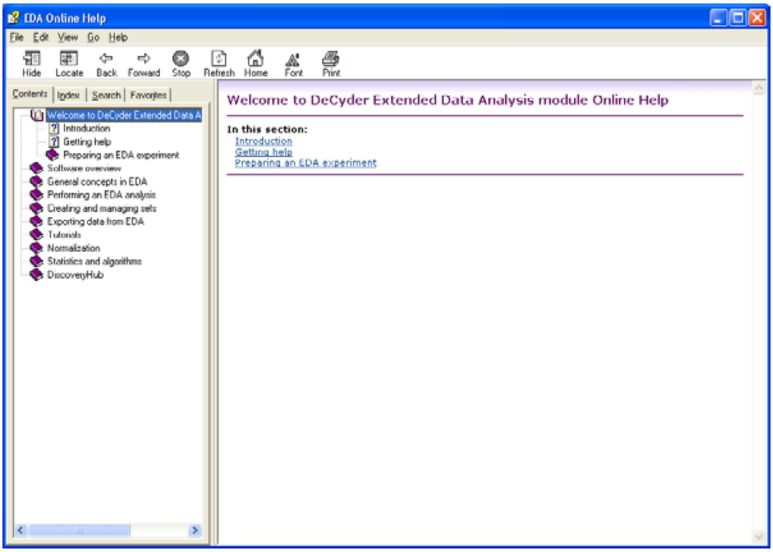
1.3 Getting help

The online help connected to the software contains more detailed information on the functionalities in EDA than the user manual. In addition to providing the same information as in the user manual, the online help also contains detailed information on the windows, dialogs and menus in EDA. This information can easily be found by using the built-in context sensitive help function, e.g. the help buttons and F1.

There are several ways to get help when using the software:

- Select **Help:Help Contents and Index...** from the menu bar to open the help file displaying the **Contents** tab and browse for help information. It is also possible to use the **Index** and **Search** tabs to search for help information.

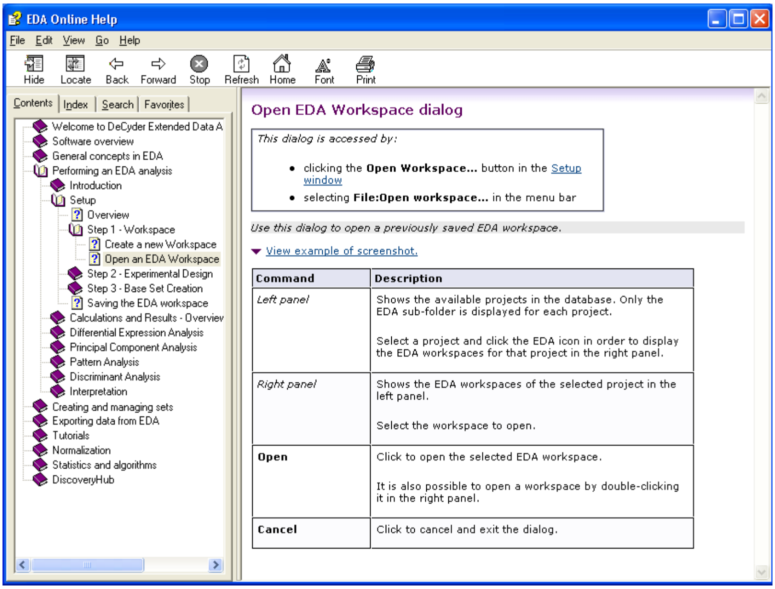

Help
Help Contents and Index...
About DeCyder Extended Data Analysis...


EDA Online Help
File Edit View Go Help
Hide Locate Back Forward Stop Refresh Home Font Print
Contents Index Search Favorites
Welcome to DeCyder Extended Data A
Introduction
Getting help
Preparing an EDA experiment
Software overview
General concepts in EDA
Performing an EDA analysis
Creating and managing sets
Exporting data from EDA
Tutorials
Normalization
Statistics and algorithms
DiscoveryHub

Welcome to DeCyder Extended Data Analysis module Online Help

In this section:
Introduction
Getting help
Preparing an EDA experiment

- Click the **Help** button in any dialog in the software to get instructions for how to enter information in that dialog.


EDA Online Help
File Edit View Go Help
Hide Locate Back Forward Stop Refresh Home Font Print
Contents Index Search Favorites
Welcome to DeCyder Extended Data A
Software overview
General concepts in EDA
Performing an EDA analysis
Introduction
Setup
Overview
Step 1 - Workspace
Create a new Workspace
Open an EDA Workspace
Step 2 - Experimental Design
Step 3 - Base Set Creation
Saving the EDA workspace
Calculations and Results - Overview
Differential Expression Analysis
Principal Component Analysis
Pattern Analysis
Discriminant Analysis
Interpretation
Creating and managing sets
Exporting data from EDA
Tutorials
Normalization
Statistics and algorithms
DiscoveryHub

Open EDA Workspace dialog

This dialog is accessed by:

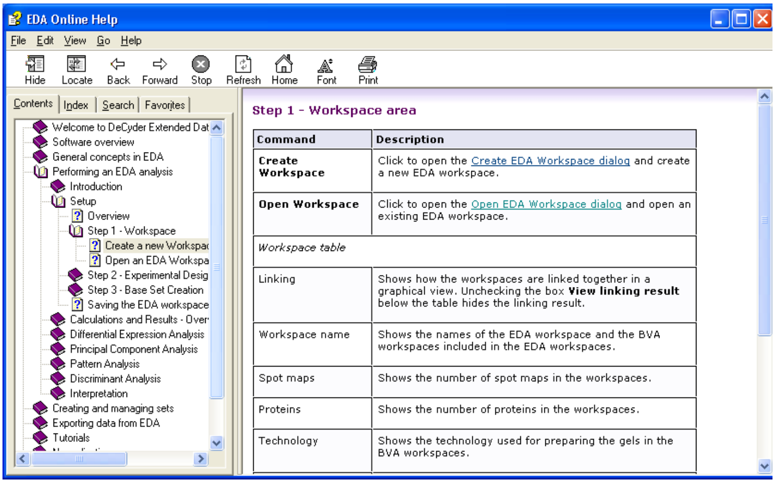
- clicking the **Open Workspace...** button in the Setup window
- selecting **File:Open workspace...** in the menu bar

Use this dialog to open a previously saved EDA workspace.

View example of screenshot.

Command	Description
Left panel	Shows the available projects in the database. Only the EDA sub-folder is displayed for each project. Select a project and click the EDA icon in order to display the EDA workspaces for that project in the right panel.
Right panel	Shows the EDA workspaces of the selected project in the left panel. Select the workspace to open.
Open	Click to open the selected EDA workspace. It is also possible to open a workspace by double-clicking it in the right panel.
Cancel	Click to cancel and exit the dialog.

- Press **F1** to open the online help for the part of the screen that is currently in focus. Focus on a graph is indicated by a thin grey border around the graph. Focus on a table is indicated by a darker color of the row and column headers. To focus on an area, click in that area.


EDA Online Help
File Edit View Go Help
Hide Locate Back Forward Stop Refresh Home Font Print
Contents Index Search Favorites
Welcome to DeCyder Extended Dat
Software overview
General concepts in EDA
Performing an EDA analysis
Introduction
Setup
Overview
Step 1 - Workspace
Create a new Workspac
Open an EDA Workspa
Step 2 - Experimental Desig
Step 3 - Base Set Creation
Saving the EDA workspace
Calculations and Results - Over
Differential Expression Analysis
Principal Component Analysis
Pattern Analysis
Discriminant Analysis
Interpretation
Creating and managing sets
Exporting data from EDA
Tutorials

Step 1 - Workspace area

Command	Description
Create Workspace	Click to open the Create EDA Workspace dialog and create a new EDA workspace.
Open Workspace	Click to open the Open EDA Workspace dialog and open an existing EDA workspace.
Workspace table	
Linking	Shows how the workspaces are linked together in a graphical view. Unchecking the box **View linking result** below the table hides the linking result.
Workspace name	Shows the names of the EDA workspace and the BVA workspaces included in the EDA workspaces.
Spot maps	Shows the number of spot maps in the workspaces.
Proteins	Shows the number of proteins in the workspaces.
Technology	Shows the technology used for preparing the gels in the BVA workspaces.

1.4 Preparing an EDA experiment

In EDA, one or several BVA workspaces can be analyzed. When setting up a new EDA experiment the BVA workspaces to include are imported into EDA. At import, the contents of the BVA workspaces that are necessary for performing analyses in EDA, are transferred from BVA to EDA (e.g. the experimental design, spot maps set to M, Master and A, Analysis etc.).If several BVA workspaces are imported, EDA will also try to link the different BVA workspaces.

To simplify the transfer of information from the BVA workspaces to EDA it is recommended to follow the guidelines in sections 1.4.1-1.4.3. The sections contain guidelines regarding the processing of gels in BVA and Batch (the step before import of the BVA workspaces into EDA).

1.4.1 General guidelines

- It is recommended to set up the experimental design for each BVA workspace in the BVA module (see section 1.4.2 for more information). However, it is also possible to set up the design in EDA.

- The gels in each BVA workspace must be matched in the BVA module. It is important to check that the matching is correct. If any statistical analyses are performed in BVA, these results will not be imported into EDA when creating the EDA workspace. The analyses should be performed in the EDA workspace in EDA.

- If possible, use the same Master for the different BVA workspaces. The single master can then be used for linking the BVA workspaces in EDA. It is also possible to use a Template for linking. See section 1.4.3 for information on how to prepare for linking of the BVA workspaces in EDA when working in BVA module or Batch module.

1.4.2 Guidelines for setting up the experimental design for BVA workspaces in the BVA module

Note: *If you want to analyze existing BVA workspaces from previous experiments, an experimental design that does not follow the guidelines below may have been assigned to the different BVA workspaces. It is then important to check the experimental design when the workspaces have been imported into EDA and, if necessary, adjust the design according to the guidelines below.*

- If the same experimental group exists in several BVA workspaces, make sure that the group has the same name in the different workspaces. When importing the BVA workspaces into EDA, the spot maps in groups with the same group name, condition names and values will be placed together in EDA.

- Make sure that all spot maps to be included in the EDA experiment have the correct Function assignment. All spot maps set to **M**, **Master** and **T**, **Template** in BVA/Batch are imported into EDA. Also, spot maps set to **A**, **Analysis** in BVA/Batch that are part of a gel where one of the other spot maps is set to Standard (i.e located in the Standard folder) are imported into EDA. The Master/Template is needed because it contains the matching information.

- If paired tests are to be performed in EDA, enter the **Sample IDs** (**Subjects** in EDA) for the different spot maps in the BVA/Batch module.

1.4.3 Preparing for the linking of BVA workspaces

Linking is performed to identify which spots are the same in the different BVA workspaces.

When importing more than one BVA workspace into EDA the BVA workspaces are automatically linked in EDA if the same Master or Template was used in the BVA workspaces.

Note: *If a common Master or Template does not exist (e.g. if the gels have been run on different pH-strips), the BVA workspaces can not be linked in EDA. However, it is still possible to analyze the workspaces in EDA.*

Linking via Master

If the same Master (i.e. the same standard) is used in the BVA workspaces, the spots on the spot maps in one BVA workspace can be linked to the corresponding spots on the spot maps in the other BVA workspace via the Master Spot number.

Only spots present on the Master are imported into EDA. Also, because the same standard was used, no normalization of the BVA workspaces needs to be performed in EDA.

Fig. 1-2 shows the linking strategy when using a common Master in the BVA workspaces. It also shows examples of reduced protein expression (indicated with a blue ring in spot map 6) and a missing value (indicated with a red ring in spot map 3) among the linked spot maps. When the data set is presented in EDA the protein expression and possible missing values are visualized in a heat map.

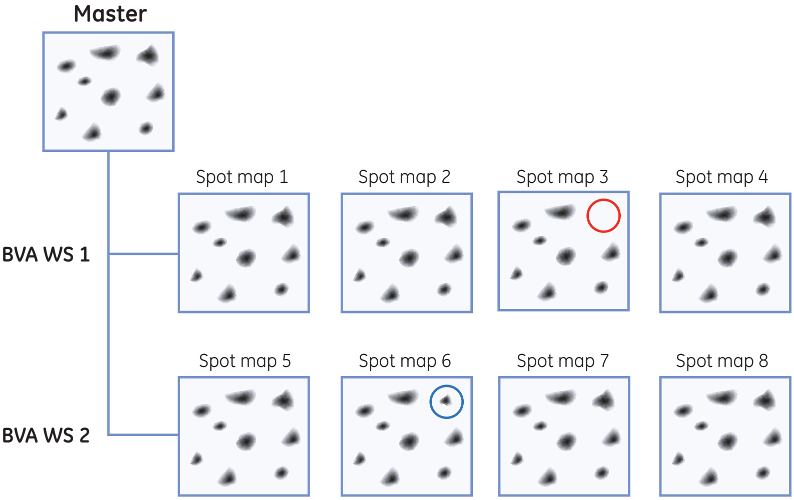


Fig 1-2. The same Master is used in both BVA workspaces. Once matching has been performed, the corresponding spots on the spot maps in BVA WS 1 and BVA WS 2 have the same Master Spot number.

Linking via Template

If different Masters are used in the BVA workspaces, linking can be performed via a Template. When linking via a Template the spots on the spot maps in one workspace can be linked to the corresponding spots on the spot maps in the other BVA workspace via the DIA spot number.

All spots present on the Masters in the BVA workspaces are transferred to EDA but only spots present on the Template spot map can be linked.

If linking via a Template, normalization of the BVA workspaces should be performed in EDA.

Note: *If linking workspaces in this way, it will not be possible to link all spots on the different spot maps. In EDA, unlinked spots will appear as missing values on the spot maps where they are not present (other BVAs).*

Fig. 1-3 shows the linking strategy when using a common Template in the BVA workspaces. It also shows examples of reduced protein expression and missing values among the linked spot maps:

- The red ring in spot map 3 indicates a missing value in BVA WS 1.
- The blue ring in spot map 6 indicates decreased protein expression.
- The red spot in Master 2 will be a missing value on the Template spot map and on all spot maps in BVA WS 1.
- The purple spot in Master 1 will be a missing value on the Template spot map and on all spot maps in BVA WS 2.

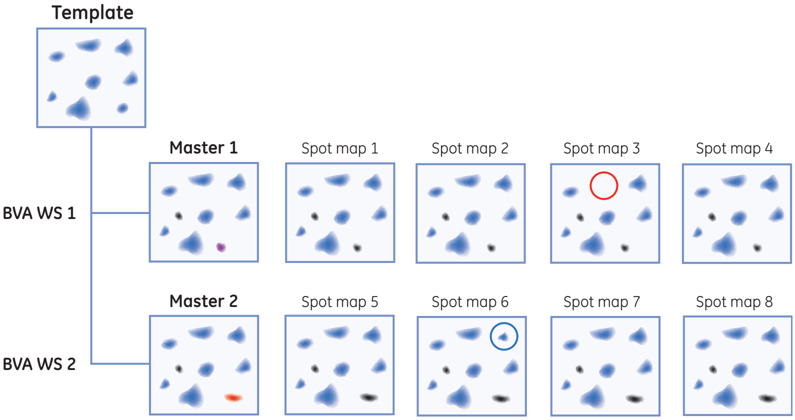


Fig 1-3. The same Template but different Masters are used for the two workspaces. The blue spots will be linked in EDA.

Prepare for linking via Master in the BVA module

1. Make sure that each BVA workspace for inclusion in the EDA workspace contains the spot map that will be used as Master.

   For information on how to add separate spot maps to a BVA workspace, see *DeCyder 2D Software Version 6.5 User Manual*.

2. For each BVA workspace to be included in the EDA workspace:

   a. Set the spot map in step 1 to **M**, **Master**.

   b. Perform matching. Once matching has been performed, the BVA workspace can be imported into EDA.

   Note: *If matching already has been performed using different Masters in the BVA workspaces, it is possible to set the spot map that will be used for linking to T, Template, instead. In this way, the matching does not need to be re-performed (which is necessary if changing Master). The BVA workspaces will be linked via the template instead. However, remember that the BVA workspaces must be normalized in EDA if linking via Template.*

Prepare for linking via Template in the BVA module

1. Make sure that each BVA workspace to be included in the EDA workspace contains the spot map (preferably a standard) that will be used as Template.

   For information on how to save a spot map as a Template and/or how to add Template spot maps to a BVA workspace, see *DeCyder 2D Software Version 6.5 User Manual*.

2. For each BVA workspace to be included in the EDA workspace:

   a. Set the spot map in step 1 to **T**, **Template**.

   b. Perform matching. Once matching has been performed, the BVA workspace can be imported into EDA.

Prepare for linking via Master in the Batch module

1. Process the DIA workspace containing the spot map (set to **M**, **Master**) that will be used as Master in all BVA workspaces.

2. Set up the batch with the other spot maps to be processed.

3. Right-click in the BVA batch list and select **Add BVA item**.

4. Add the processed DIA workspace (from step 1), containing the spot map to be used as Master, to the new batch.

5 Set this spot map to Master (**M**).

6 Run the batch. Once the batch has been run, it is possible to import the BVA workspaces into EDA.

Prepare for linking via Template in the Batch module

1 Make sure that the spot map to be used as the template has been saved as a Template.

For information on how to save a spot map as a Template, see *DeCyder 2D Software Version 6.5 User Manual*.

2 Set up one batch with all spot maps.

3 Right-click in the BVA batch list and select **Add BVA item**.

4 Locate the appropriate folder named **BVA Templates** and add the template to the BVA workspace.

5 In the Batch module, set this spot map to **Template**, **T**.

6 Run the batch. Once the batch has been run, it is possible to import the BVA workspaces into EDA.

2 Software overview

2.1 Computer requirements and database administration

Please refer to *DeCyder 2D Software Version 6.5 User Manual* for information on computer requirements and database administration.

2.2 Structure of the EDA part of DeCyder 2D Software

The structure of the EDA part of DeCyder 2D Software is outlined below. For a complete description of the structure of DeCyder 2D Software, see *DeCyder 2D Software Version 6.5 User Manual*.

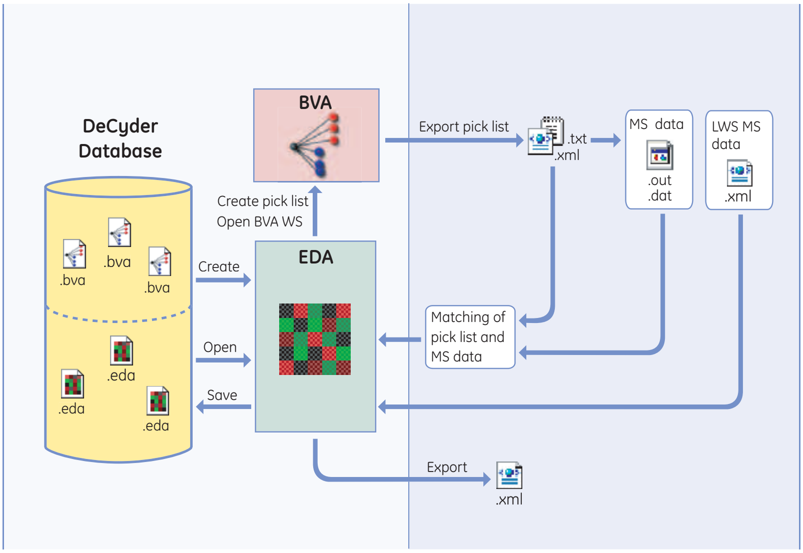
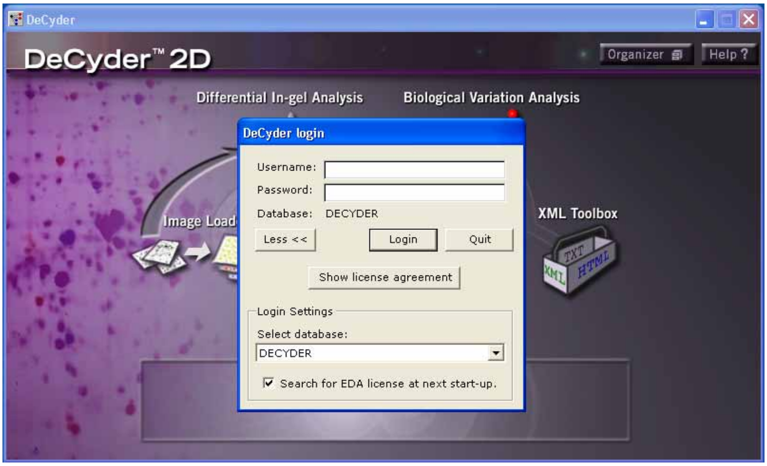


Fig 2-1. Structure of the EDA part of DeCyder 2D Software.

2.3 Start DeCyder 2D 6.5 Software

1 Select **Start:All Programs:DeCyder™ 2D 6.5 Software:DeCyder™ 2D 6.5**. *Alternatively*, double-click the DeCyder icon on the desktop.

The **DeCyder™ 2D** start screen and the **DeCyder Login** dialog will open.



DeCyder login

Username:

Password:

Database: DECYDER

Login Settings

Select database: DECYDER

☑ Search for EDA license at next start-up.

2 It is possible to view the license agreement by clicking the **Show license agreement** button.

3 Make sure that the box **Search for EDA license at next start up** is checked.

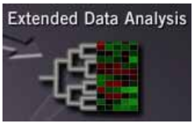
*If the **Search for EDA license at next start up** box is unchecked*: check it, click **Quit** and re-start the software.

4 Make sure that the correct database is selected in the **Select database** field, otherwise select as appropriate.

5 Enter **User name** and **Password** and click **Login** to log into the software. The DeCyder start screen is activated.

Note: *The license files for EDA must be able to be located for the EDA module to appear in the DeCyder™ 2D start screen.*

2.4 Open EDA



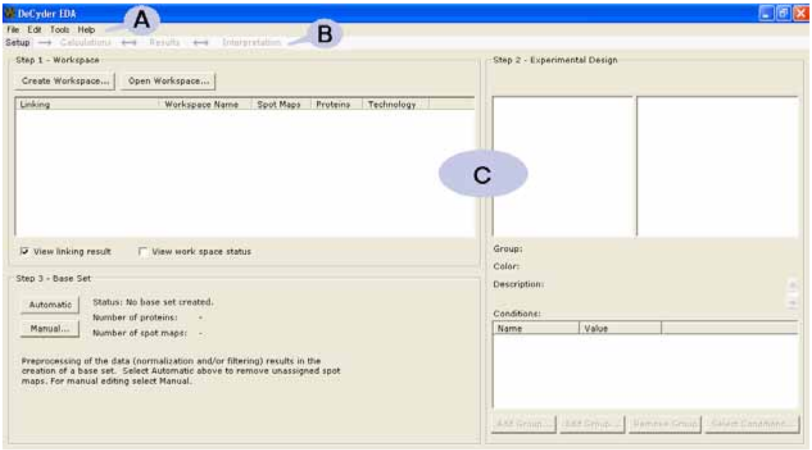
After logging into the DeCyder 2D Software, click the Extended Data Analysis (EDA) icon in the DeCyder 2D main window. EDA will open displaying the DeCyder EDA main screen.

2.5 DeCyder EDA main screen

The DeCyder EDA main screen is divided into three areas:

- menu bar (**A**)
- workflow area (**B**)
- work area (**C**)

Depending on the currently selected step in the workflow area, the work area will appear different. In the beginning, the first step in the workflow, **Setup**, is selected and the **Setup** window is displayed in the work area.



Status: No base set created.
Number of proteins: -
Number of spot maps: -

Preprocessing of the data (normalization and/or filtering) results in the creation of a base set. Select Automatic above to remove unassigned spot maps. For manual editing select Manual.

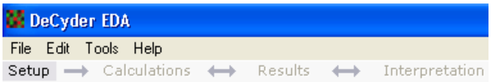
2.5.1 Menu bar

The menu bar contains 4 different drop-down menus, i.e. **File**, **Edit**, **Tools** and **Help**. These are used to create, open, save and export workspaces, create pick lists, import mass spectrometry (MS) Data, open BVA source files in the BVA module, copying data, managing sets and web links and viewing the online help.

For a description of the commands in the different menus, see the online help.

2.5.2 Workflow area

The software consists of 4 major workflow steps displayed at the top. When clicking on a step, a corresponding window opens in the work area. In the beginning, only the Setup step is available and the other steps that are not available have a dimmed appearance. All steps will be available once the base set has been created.



The different steps in the workflow area are listed with a short description in the Table 2-1. For information on how to perform an EDA analysis see Chapter 4, Performing an EDA analysis - Introduction.

Table 2-1. Description of the main steps in the workflow area.

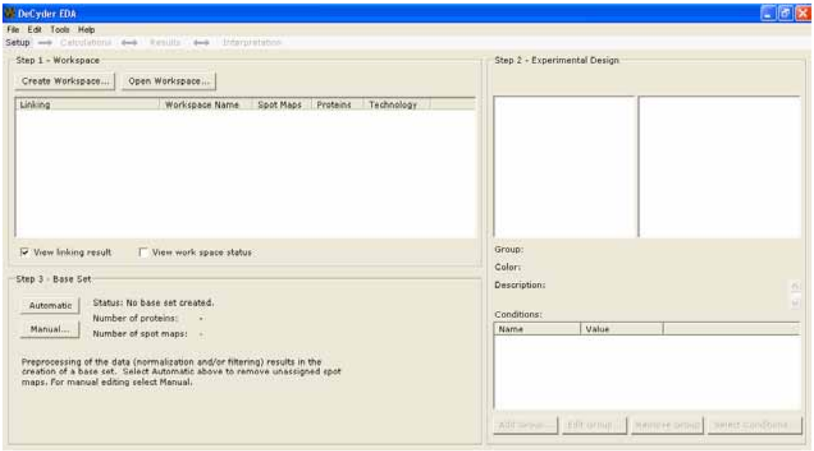
Workflow step	Description
Setup	In this step, the EDA workspace is set up and a base set created. For more information on setup, see Chapter 5.
Calculations	In this step, the calculations to be performed are set up and calculated. Several calculations can be set up and performed one after the other. For more information, see Chapter 6.
Results	In this step, the results of the performed calculations can be viewed and analyzed. Once the analyses have been performed it is possible to perform new calculations, biological interpretation or export the results. For more information, see Chapter 6.
Interpretation	In this step, biological interpretation of the selected proteins are performed by integrating biological information and context from in-house or public databases. For more information, see Chapter 11.

2.5.3 Work area

The work area displays different windows depending on the position in the workflow area.

When creating a new workspace, the **Setup** window is displayed.

See Chapter 4-11 for information on how to enter settings in the different windows that appear in the work area.



Status: No base set created.
Number of proteins: -
Number of spot maps: -

Preprocessing of the data (normalization and/or filtering) results in the creation of a base set. Select Automatic above to remove unassigned spot maps. For manual editing select Manual.

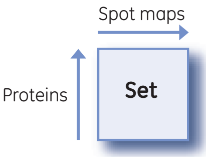
2.6 DeCyder EDA Software keyboard shortcuts

Shortcut	**Description**
Workspace	
Alt + F	Open File menu
Alt + E	Open Edit menu
Alt + T	Open Tools menu
Alt + H	Open Help menu
File menu	
Ctrl + N	Create new EDA workspace
Ctrl + O	Open EDA workspace
Ctrl + S	Save workspace
Edit menu	
Ctrl + C	Copy
Tools menu	
Ctrl + M	Show Manage sets dialog
Ctrl + U	Opens the BVA workspace displaying the spot and spot map selected in EDA.

3 General concepts in EDA

3.1 The set concept

EDA uses a set of data for the analysis. A set is a group of spot maps (or experimental groups) with matched spots, i.e. a group of spot maps and proteins.



A set of data can be displayed in several ways depending on the context.

For example, when filtering a set to remove proteins and/or spot maps, a heat map is displayed, showing an overview of the data (see Fig. 3-1). See section 3.3 for more information on the heat map.

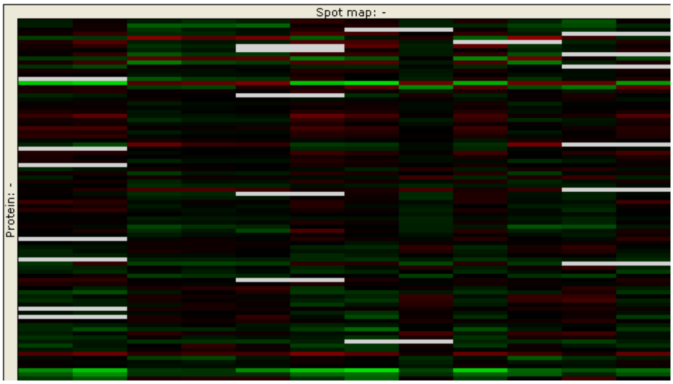


Fig 3-1. Heat map showing an overview of the data set.

The results of the statistical analyses that can be set up in the software will also be presented in such a way that selection of interesting proteins/spot maps is facilitated, e.g. by grouping/ordering of proteins/spot maps. Depending on the analysis, the data set can be presented in tables and/or heat maps or in different kinds of graphs and plots.

3.2 Working with sets

There are different levels of sets (see Fig. 3-2). When working in EDA a base set must be created from the original data set on which further analysis will be performed. New sets can then be created and combined in various ways.

The ***original data set*** consists of the BVA workspaces imported and linked in the EDA workspace.

Note: *Only standard spot maps set to M, Master, and non-standard spot maps set to A, Analysis, containing standard abundance values in BVA are imported into EDA. For example, a pick gel set to A and P, Pick, is not imported unless it contains standard abundance values.*

Filtering and normalization of the data is then performed to create a ***base set*** (see section 5.4 for information on how to create the base set) on which statistical calculations can be performed.

Based on the results of the calculations, new ***sets*** can be created (all sets will be displayed in the software in a drop-down list) and new calculations/biological interpretation can be performed.

Created sets can also be ***combined*** in various ways by using the logical conditions **AND** and **OR** to create a new set (see section 12.2, Managing sets for more information).

New sets can be created and calculations can be performed until you are satisfied with the results.

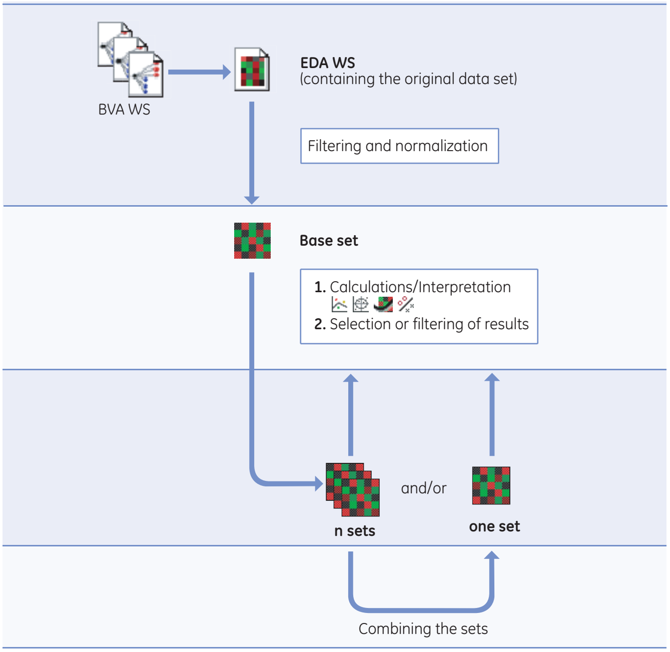

BVA WS
EDA WS
(containing the original data set)
Filtering and normalization
Base set
1. Calculations/Interpretation
2. Selection or filtering of results
n sets
and/or
one set
Combining the sets

Fig 3-2. Set concept.

Note: *The original data included in the EDA analysis are not changed during the analysis. Creating sets only helps to view selected data and to perform statistical analyses only on the data in that set.*

3.3 The heat map

The heat map can be likened to a coordinate system with proteins on the y-axis and spot maps/experimental groups on the x-axis. Each coordinate shows:

- the expression of a protein (log standard abundance) on a spot map (if showing proteins and spot maps)

  or

- the mean of a protein's expression (log standard abundance) on all spot maps in an experimental group (if showing proteins and experimental groups)

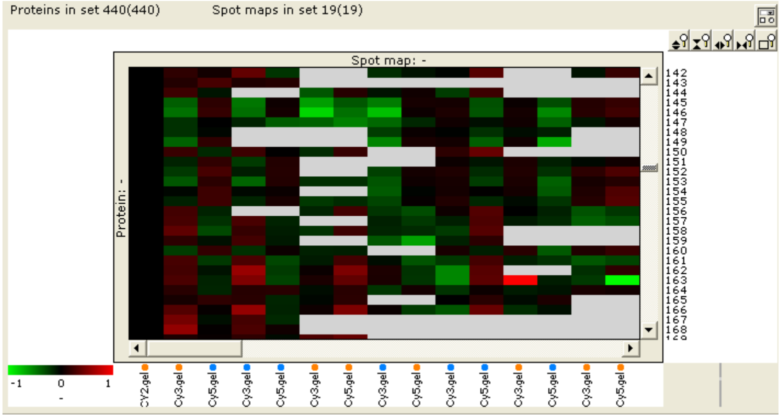
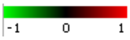


Fig 3-3. Enlarged part of a heat map.



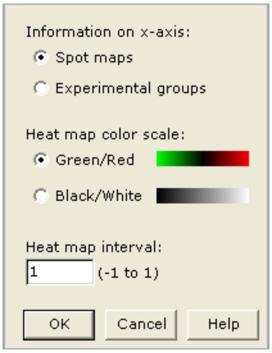
Each coordinate displays a color representing the protein expression. By default the color scale goes from ***green*** (decreased protein expression) to ***black*** (no change in protein expression) to ***red*** (increased protein expression) and is displayed at the bottom-left corner.

If no data exists for a coordinate (***missing value***) this coordinate is displayed in gray. Below the color scale, the log standard abundance value interval for the colors are displayed. The interval is set to -1 to 1 by default.

3.3.1 Changing the heat map settings

It is possible to change the information on the x-axis (display of spot maps or experimental groups), the heat map color scale (color or gray scale) and the heat map interval.

Click the **Settings** icon to display the **Heat map settings** pop-up dialog and change the settings as appropriate.



Information on x-axis:

● Spot maps

○ Experimental groups

Heat map color scale:

● Green/Red

○ Black/White

Heat map interval:

1 (-1 to 1)

Fig 3-4. Heat map settings pop-up dialog.

Table 3-1 summarizes the settings for the heat map color/gray scale and heat map interval.

Table 3-1. Summary of heat map color scale and heat map interval settings.

Color scale Green/Red	Gray scale Black/White	Protein expression	Interval value examples
Green	Black	Decreased	-2 means 100 fold decrease -1 means 10 fold decrease
Black	Gray	Unchanged	0 means no change in protein expression
Red	White	Increased	1 means 10 fold increase 2 means 100 fold increase
Gray	Pink	Missing value	-

3.3.2 Zooming within the heat map

It is possible to zoom in the heat map and graphs using the zooming bar displayed at the top-right corner of the heat map/graph. The table below lists descriptions of the different zooming options.

Icon	Description
⬍	Use to zoom in vertically.
⧗	Use to zoom out vertically.
⬌	Use to zoom in horizontally.
⧓	Use to zoom out horizontally.
☐	Use to fit the graph/heat map to the window.

To zoom in the heat map using the mouse:

1 Click the mouse button where the upper left corner of the heat map is to be located, and drag the pointer to where the lower right corner is to be located.

A rectangle appears in the heat map.

2 Release the mouse button. The heat map immediately zooms to the area specified.

To zoom out of the heat map using the mouse:

1 Click the mouse button on the heat map and drag the pointer up and left.

A rectangle appears in the heat map.

2 Release the mouse button. The heat map immediately zooms out. The heat map is always fitted to the window.

4 Performing an EDA analysis - Introduction

4.1 Steps involved in analysis using EDA

Before any analyses can be performed in EDA, spot detection (in DIA module, giving a log standard abundance value for each spot) and matching of the spot maps (in BVA module) must have been performed. See the *DeCyder 2D Software Version 6.5 User Manual* for information on spot detection and matching.

In EDA, statistical analyses are performed using a number of complex algorithms. The analyses in EDA include four main workflow steps:
Setup, **Calculations**, **Results** and **Interpretation**.

The algorithms associated with these steps form part of the in-built functionality of the DeCyder EDA module. See Table 4-1 for information on the different steps.

Workflow step	Description
Setup	This step includes setting up the EDA workspace, defining/ checking the experimental design and pre-processing the data to create a base set.
Calculations	This step includes choosing a set on which to perform calculations, setting up the statistical analyses to be performed, adding the analyses to the calculation list, and calculating them. Depending on the biological queries, different calculations can be performed. ***Tip!*** *Work through the tutorials for examples of possible workflows.*
Results	This step includes analyzing the results of the performed calculations. It is possible to create new sets containing proteins of interest (which are to be further studied) and go back to the **Calculations** step and perform new calculations on the new set, or go to the **Interpretation** step and perform biological interpretation. It is also possible to go back to the **Calculations** step, change settings and perform re-calculations on the old set.
Interpretation	This step includes the biological interpretation of the selected proteins by integrating biological information and context from in-house or public databases. MS data must be available in order to perform the analysis. ***Note:*** *It is possible to generate pick lists from EDA and import MS data into EDA.*

Table 4-1. An overview of the main workflow steps in EDA.

When satisfied with the analyses, data can be exported from EDA (see Chapter 13, Exporting data from EDA).

Fig. 4-1 outlines an overview of the workflow in EDA independent of any biological queries.

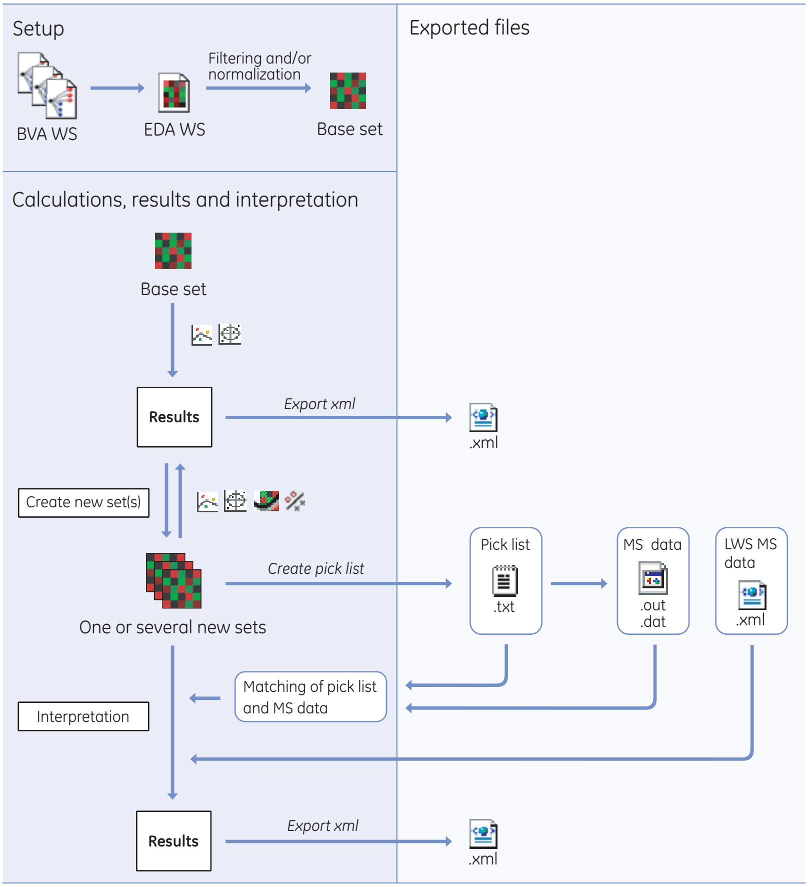


Fig 4-1. An overview of the workflow in EDA.

Note: *Depending on the biological queries, the workflow of which calculations and analyses to perform will change. The two tutorials in Chapter 15 and 16 gives examples of how the different statistical methods in the software can be used and describes the most frequently used methods.*

5 Setup

5.1 Overview

The first step to perform in EDA before any analysis can be made is to set up the EDA workspace and create a base set of the original data (BVA workspaces). Setup includes importing the BVA workspaces to be included in the EDA workspace, defining/checking the experimental design and creating a base set. All of this is performed in the **Setup** window which consists of 3 main steps:

- **Step 1 - Workspace (A)**
  This step includes creating an EDA workspace by importing BVA workspaces. Information on the imported workspaces and how they are linked together are displayed.

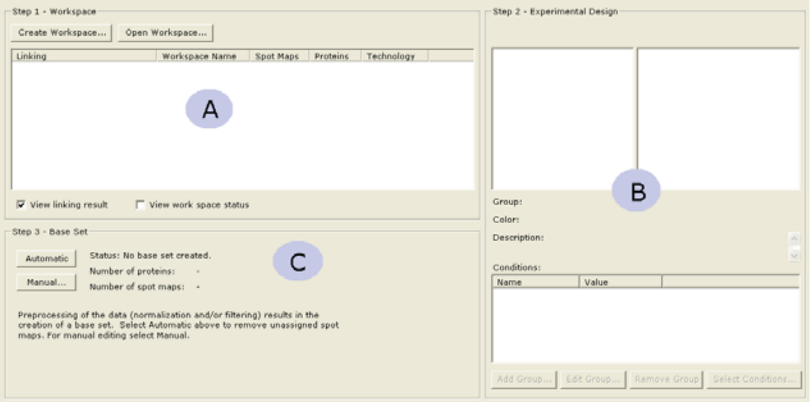
  See section 5.2 for more information.

- **Step 2 - Experimental Design** (**B**)
  This step includes assigning experimental groups and conditions for the different samples included in the EDA workspace.

  See section 5.3 for more information.

- **Step 3 - Base Set Creation** (**C**)
  This step includes creating the base set automatically or manually by filtering and normalization of the data.

  See section 5.4 for more information.



Status: No base set created.
Number of proteins: -
Number of spot maps: -

Preprocessing of the data (normalization and/or filtering) results in the creation of a base set. Select Automatic above to remove unassigned spot maps. For manual editing select Manual.

Fig 5-1. Setup window.

5.2 Step 1 - Workspace

The first step in the setup is to create an EDA workspace including one or several BVA workspaces. It is also possible to open an already existing EDA workspace (see section 5.2.2, Open an EDA Workspace).

When creating an EDA workspace the BVA workspaces are imported into the EDA workspace. Not all information saved in the BVA workspaces or all of the spot maps are imported into EDA. The following content of the BVA workspaces are imported:

- *Standard abundance values*
  The standard abundance values are imported into EDA. These values are the only data used when performing statistical analyses.

- *Spot maps set to **M**, **Master** and **T**, **Template** in BVA*
  Only spot maps set to **M**, **Master**, and **T**, **Template** in BVA are imported into EDA.

  Note: *Standards that are not set to Master or Template are not needed in EDA. In BVA, the standards are needed for matching and calculation of the standard abundance values (which are the only data used when performing statistical analyses). Because the standard abundance values are imported into EDA there is no longer any need for the standards except those set as **Master**/**Template** (containing the matching information).*

- *Non-standard spot maps set to **A**, **Analysis** that are also part of a gel where one of the other spot maps are set to Standard*
  All spot maps set to **A**, **Analysis**, in BVA that are also part of a gel where one of the other spot maps are set to Standard (located in the Standard folder) are imported into EDA.

- *Experimental design*
  The experimental design in the BVA workspaces is transferred to EDA. See section 5.3 for more information on the experimental design.

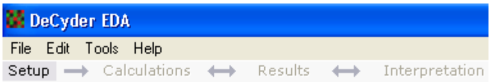
- *MS Data, Sample IDs*
  Available MS Data and Sample IDs in BVA are imported into EDA.

Note: *No statistical values are copied into the EDA workspace. All statistical analyses available in BVA can be performed in EDA.*

5.2.1 Create a new Workspace

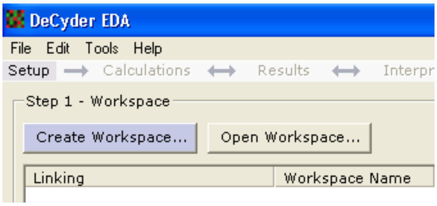
To create a new workspace:

1 In the workflow area, click **Setup**.

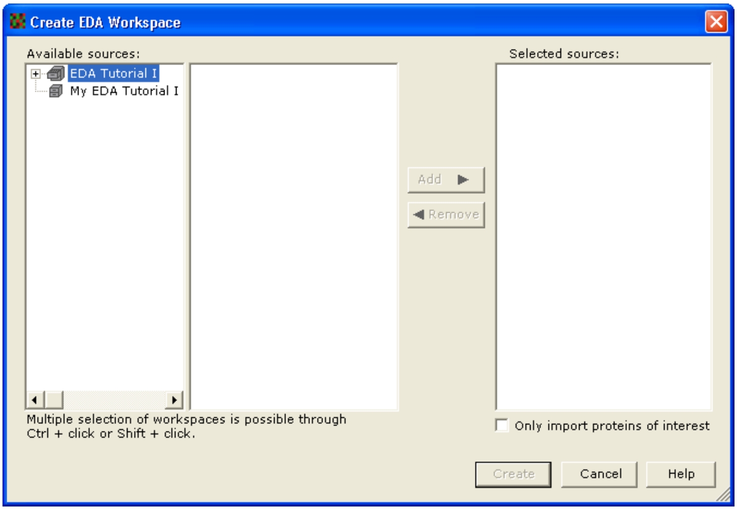


The **Setup** window is displayed in the work area.

2 Click **Create Workspace...** in the **Step 1 - Workspace** area.



The **Create EDA Workspace** dialog is displayed.



Multiple selection of workspaces is possible through Ctrl + click or Shift + click.

3 Double-click a project in the **Available Workspace(s)** area (in the left panel) and click the BVA folder. The BVA workspaces included in the project are shown to the right.

4 Select the BVA workspaces to include in the EDA workspace and click **Add -->**.

Added BVA workspaces are displayed in the **Selected sources** area (right panel).

To add more BVA workspaces, select the appropriate workspace and click **Add -->** again. To remove workspaces from the EDA workspace, select the appropriate workspace in the right panel and click **<-- Remove**.

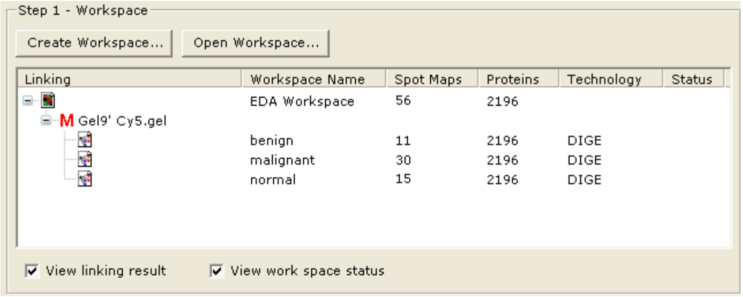
5 Repeat step 3 and 4 until all BVA workspaces to be included in the EDA workspace are listed in the **EDA Workspace** area.

6 If you only want to import spots set to proteins of interest in BVA, check the **Only import proteins of interest** box.

7 When all required BVA workspaces have been added, click **Create** to create the EDA workspace. The BVA workspaces are copied into the EDA workspace.

During import, the software searches for a common Master or Template (in that order), to link the BVA workspaces together.

8 The created EDA workspace with links (if available) will be displayed in the **Step 1 - Workspace** area of the **Setup** window.



Linking	Workspace Name	Spot Maps	Proteins	Technology	Status
	EDA Workspace	56	2196		
M Gel9' Cy5.gel					
	benign	11	2196	DIGE	
	malignant	30	2196	DIGE	
	normal	15	2196	DIGE	

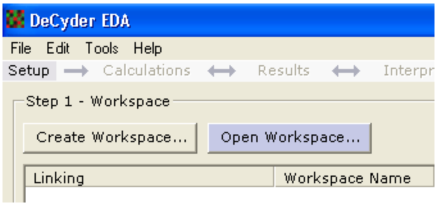
☑ View linking result ☑ View work space status

An **M** in the Linking column means that the BVA workspaces are linked by a common Master and a **T** that the BVA workspaces are linked by a template spot map.

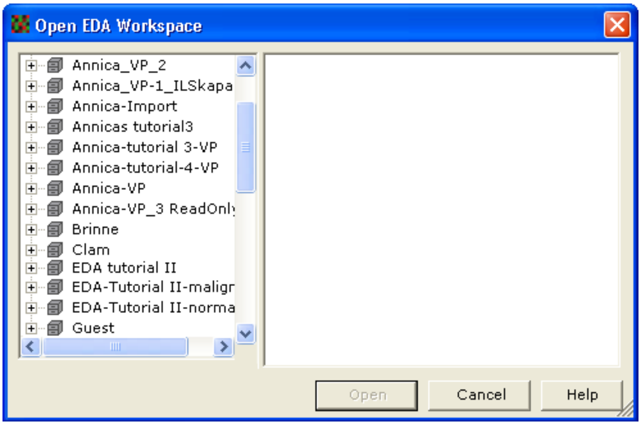
5.2.2 Open an EDA Workspace

To open previously created and saved EDA workspaces:

1 Click **Open Workspace...** in the **Step 1 - Workspace** area of the **Setup** window.



The **Open EDA Workspace** dialogis displayed.



2 Select the project (in the left panel) and then locate the **EDA workspace** file (in the right panel) to be opened. When the file is located, double-click the **EDA workspace** file *or* select the file and click **Open**.

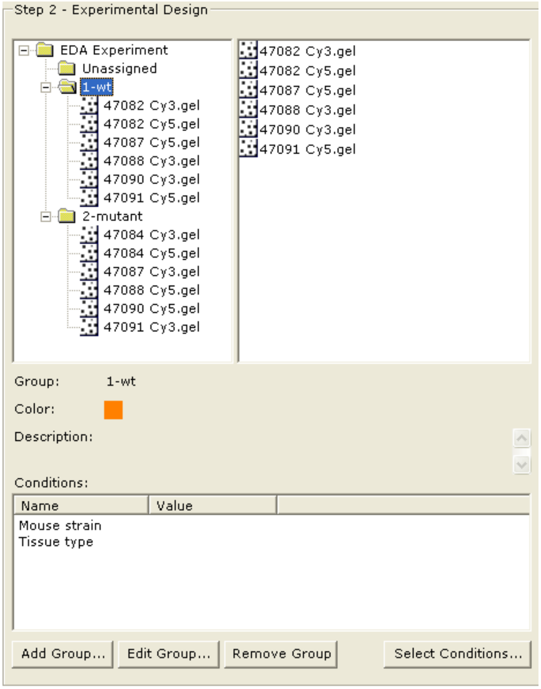
5.3 Step 2 - Experimental Design

The second step in the setup is to define or check and if necessary adjust the experimental design for the experiments.

- If an experimental design has already been defined for each BVA workspace in the BVA module, check that the design is correct. See section 5.3.1 for more information.
- If no experimental design has been defined for one or several BVA workspaces, set up the design in EDA. See section 5.3.2 for more information.

5.3.1 Check the experimental design

In the BVA module an experimental design may already have been defined for each BVA workspace. When including the BVA workspaces in the EDA workspace, the design for the BVA workspaces is transferred to the EDA workspace (see Fig. 5-2).



Step 2 - Experimental Design

EDA Experiment
Unassigned
1-wt
47082 Cy3.gel
47082 Cy5.gel
47087 Cy5.gel
47088 Cy3.gel
47090 Cy3.gel
47091 Cy5.gel
2-mutant
47084 Cy3.gel
47084 Cy5.gel
47087 Cy3.gel
47088 Cy5.gel
47090 Cy5.gel
47091 Cy3.gel

47082 Cy3.gel
47082 Cy5.gel
47087 Cy5.gel
47088 Cy3.gel
47090 Cy3.gel
47091 Cy5.gel

Group: 1-wt
Color: 🟧
Description:

Conditions:

Name	Value
Mouse strain	
Tissue type	

Fig 5-2. Imported BVA Workspaces with experimental design transferred.

Check the following:

- Make sure that the spot maps belonging to the same experimental groups in the different BVA workspaces have been placed in one group in EDA.

  Master spot maps and spot maps that were located in the Unassigned group in the BVA workspaces will be placed in the Unassigned group in EDA. These spot maps may need to be assigned to a group, see section 5.3.5.

- Check that the experimental groups have different colors by clicking on one group at a time to view the color for that group in the **Color** field. Different colors on the experimental groups facilitate the analyses in EDA.

  To edit the color for a group, select the group and click **Edit group...**. See section 5.3.6 for more information.

- Make sure that the same conditions have the same name.

- Add more conditions to the workspace, if required (see section 5.3.3 for more information).

  Note: *In EDA workspaces up to 15 conditions can be defined and either text or numerical values can be entered compared to two numerical conditions in BVA.*

When the experimental design has been checked, proceed with section 5.4, Step 3 - Base Set Creation.

5.3.2 Define an experimental design

If no experimental design has been defined for the BVA workspaces, set up the design in EDA. All imported spot maps are located in the **Unassigned** group at the beginning.

To set up the experimental design:

1. Add the conditions to be used in the experiment to the EDA workspace. If required, create new conditions. See section 5.3.3, Select conditions for a workspace.

2. Add the appropriate experimental groups. For each experimental group; assign a color and enter condition values. See section 5.3.4, Add experimental groups.

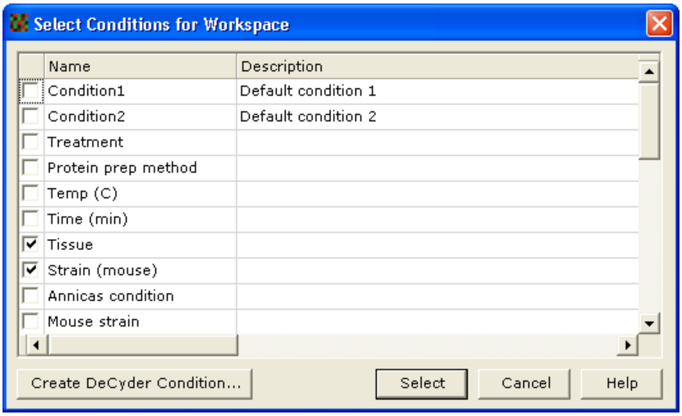
3. Assign the spot maps from the **Unassigned** group to the correct experimental group. See section 5.3.5, Assign spot maps to experimental groups.

5.3.3 Select conditions for a workspace

In BVA workspaces two numerical conditions can be defined. In EDA workspaces up to 15 conditions can be defined and either text or numerical values can be entered. The defined conditions are displayed in the **Conditions** list.

To select conditions for the workspace:

1 Click the **Select Conditions...** button in the **Step 2- Experimental Design** area. The **Select Conditions for Workspace** dialog is displayed listing the available conditions in the database.

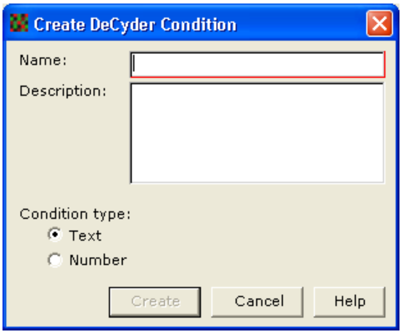


	Name	Description
□	Condition1	Default condition 1
□	Condition2	Default condition 2
□	Treatment	
□	Protein prep method	
□	Temp (C)	
□	Time (min)	
☑	Tissue	
☑	Strain (mouse)	
□	Annicas condition	
□	Mouse strain	

2 Select a condition to be included in the EDA workspace by checking the appropriate box.

Alternatively, if the required condition is not available in the list:

a. Click **Create DeCyder Condition...**. The **Create DeCyder Condition** dialog is displayed.



Name:

Description:

Condition type:
● Text
○ Number

b. Enter a **Name** for the condition and, if required, a **Description** of the condition.

c. Select whether the **Condition type** should be **Text** or **Number** by choosing the appropriate radio button.

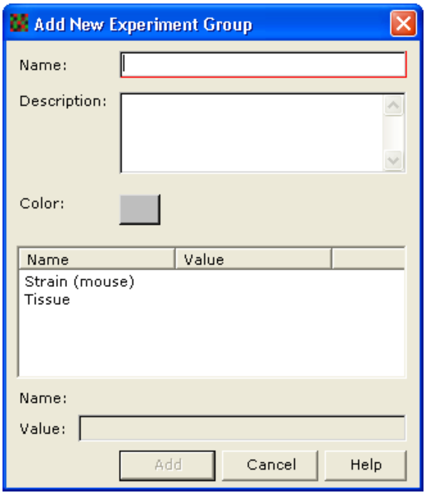
d. Click **Create** to create the condition and return to the **Select Conditions for Workspace** dialog.

3 Click **Select** to include the condition in the workspace.

5.3.4 Add experimental groups

To add a new group:

1 Click **Add group...** in the **Step 2- Experimental Design** area. The **Add New Experiment Group** dialog is displayed.



Name	Value
Strain (mouse)	
Tissue	

2 Enter a **Name** for the experimental group.

3 Enter a **Description** of the experimental group (optional) and, if required, change the color of the group by clicking the colored button next to the **Color** field.

Note: *Make sure that the color assigned to the group is unique for that group. Different colors on the experimental groups facilitate the analyses of the results.*

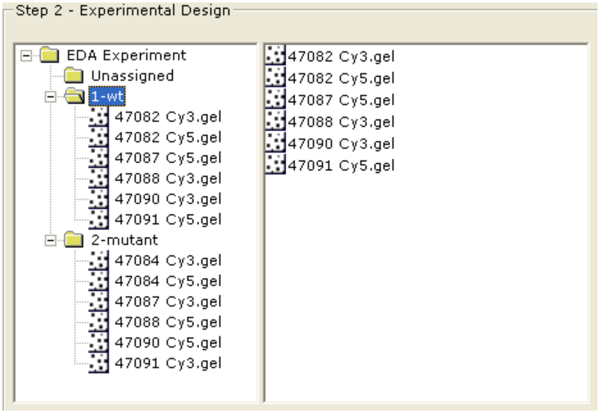
4 Select a **Condition** in the table and enter a value (numerical or text) for the condition in the **Value** field.

5 Click **Add** to add the new group to the **Experimental Design** list and close the dialog.

5.3.5 Assign spot maps to experimental groups

1 Click the **Unassigned** folder on the left to display the contents of the folder both in the left and right panels of the **Step 2 - Experimental Design** area.

If moving a spot map from one experimental group to another, select that group.


Step 2 - Experimental Design

```
EDA Experiment
    Unassigned
    1-wt
        47082 Cy3.gel
        47082 Cy5.gel
        47087 Cy5.gel
        47088 Cy3.gel
        47090 Cy3.gel
        47091 Cy5.gel
    2-mutant
        47084 Cy3.gel
        47084 Cy5.gel
        47087 Cy3.gel
        47088 Cy5.gel
        47090 Cy5.gel
        47091 Cy3.gel
```

```
47082 Cy3.gel
47082 Cy5.gel
47087 Cy5.gel
47088 Cy3.gel
47090 Cy3.gel
47091 Cy5.gel
```

2 In the right panel, select the spot map(s) to be assigned to a group and drag-and-drop the spot map(s) in that group to the left panel.

Note: *Several spot maps can be selected by pressing the Ctrl or Shift keys and clicking the spot maps.*

If a new group needs to be added, see section 5.3.4, Add experimental groups.

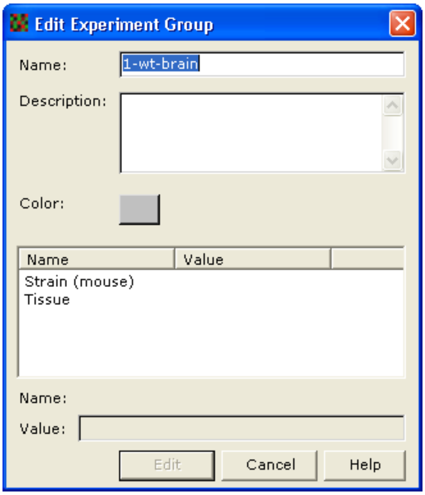
3 The spot map(s) will appear in the group to which it was dragged-and-dropped.

5.3.6 Edit experimental groups

To edit a group and its condition values:

Edit Group...

1 Click **Edit group...** in the **Step 2- Experimental Design** area. The **Edit Experiment Group** dialog is displayed.


Edit Experiment Group

Name: 1-wt-brain

Description:

Color:

Name	Value
Strain (mouse)	
Tissue	

Name:

Value:

2 Edit the information as required and click **Edit** (see section 5.3.4, Add experimental groups for information on settings).

Note: *Make sure that the color assigned to the group is unique for that group. Different colors on the experimental groups facilitate the analyses of the results.*

5.3.7 Remove experimental groups

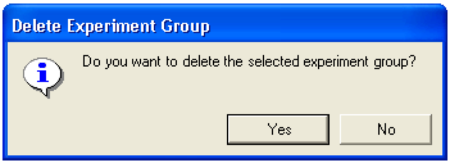
To remove a group:

1 Select the group to be removed in the left panel of the **Step 2 - Experimental Design** area.

Remove Group

2 Click **Remove group** at the bottom of the area.

A dialog appears asking you to confirm the removal of the group.


Delete Experiment Group

Do you want to delete the selected experiment group?

3 Click **Yes** to remove the group.

5.4 Step 3 - Base Set Creation

The third step in the Setup is to create the base set in the EDA workspace. This includes protein and spot map filtering and possibly normalization of data.

When creating the base set it is recommended to remove all unassigned spot maps and to remove spots with too many missing values. For example, keep spots that are present in at least 75% of the spot maps and remove all other spots. If too many missing values are present for a spot, this will affect the results of the analyses in EDA (mainly PCA).

Normalization of the data in EDA should be performed only if *two or more BVA workspaces that do not use the same standard* are included in the EDA workspace.

The base set can be created either automatically (using default values) or manually.

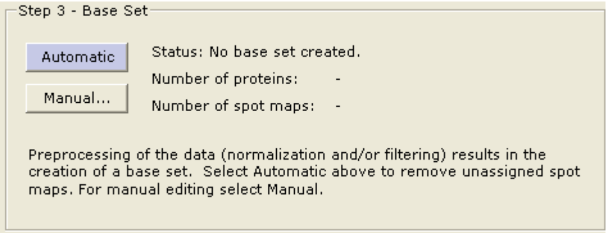
5.4.1 Create the base set automatically

When creating the base set automatically, only unassigned spot maps are removed from the original data set. No other filters are applied.

Note: *Create the base set manually if other filters (such as removing spots with too many missing values) should be applied to the data or if normalization should be performed. See section 5.4.2, Create the base set manually.*

To create a base set automatically:

1 Click **Automatic** in the **Step 3 - Base Set Creation** area.



Status: No base set created.
Number of proteins: -
Number of spot maps: -

Preprocessing of the data (normalization and/or filtering) results in the creation of a base set. Select Automatic above to remove unassigned spot maps. For manual editing select Manual.

2 The base set is created. During the creation the status **Creating base set** is displayed in the status field.

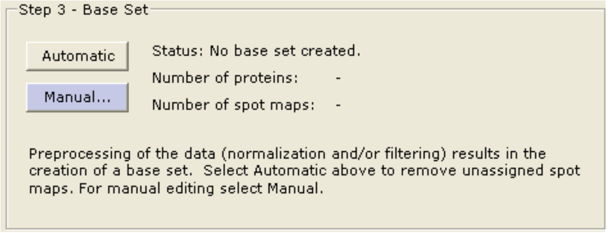
3 When the base set has been created, the status **Base set created-Calculation is now possible** is displayed in the **Status** field and the number of proteins and spot maps included in the base set are listed. Proceed with section 5.5, Saving the EDA workspace.

5.4.2 Create the base set manually

Create the base set manually to apply protein and spot map filters to the data and/or if normalization should be performed.

To create a base set manually:

1 Click **Manual...** in the **Step 3 - Base Set Creation** area.



Step 3 - Base Set

Automatic

Manual...

Status: No base set created.

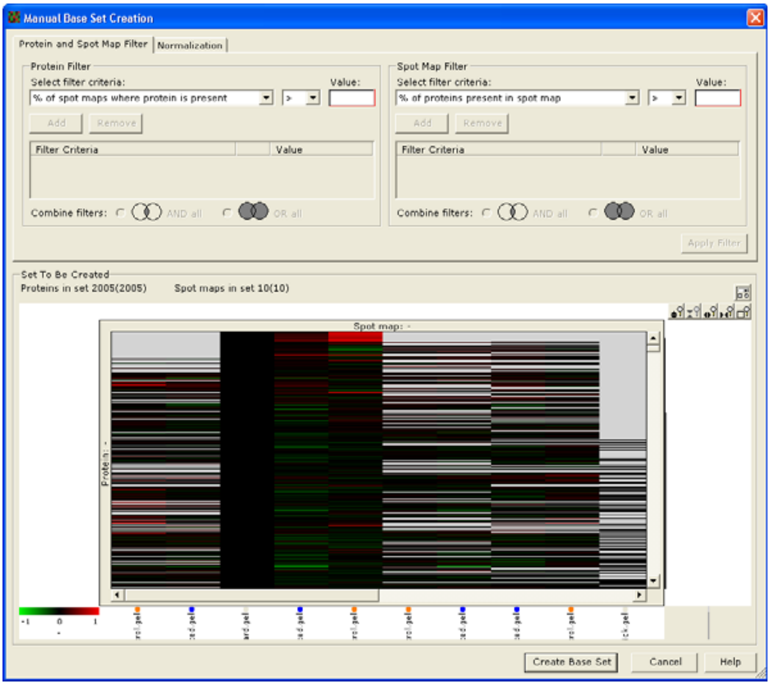
Number of proteins: -

Number of spot maps: -

Preprocessing of the data (normalization and/or filtering) results in the creation of a base set. Select Automatic above to remove unassigned spot maps. For manual editing select Manual.

The **Manual Base Set Creation** dialog opens displaying the **Protein and Spot Map Filter** tab by default.

The data set is displayed in the form of a heat map in the **Set To Be Created** area. This area also displays the number of proteins and spot maps currently included in the data set.
For more information on the heat map, how to change settings and how to zoom in the heat map, see section 3.3.



Manual Base Set Creation

Protein and Spot Map Filter | Normalization

Protein Filter

Select filter criteria: % of spot maps where protein is present; >; Value:

Spot Map Filter

Select filter criteria: % of proteins present in spot map; >; Value:

Set To Be Created

Proteins in set 2005(2005)

Spot maps in set 10(10)

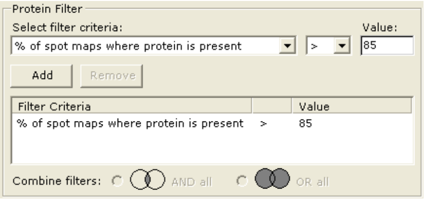
2 *Define Protein Filter criteria:*

a. Select filter criteria and one of the operators **<**, **<=**, **=**, **>=** or **>** from the drop-down lists in the **Select filter criteria** field. For information on the different criteria, see section 5.4.3, Protein and spot map filter criteria.

b. Enter a value for the criteria in the **Value** field and click **Add** to add the filter criteria to the list. If required, repeat step 2a and 2b to add a new protein filter criteria to the list.

c. Combine the **Protein Filter** criteria for the protein filter by using the logical conditions **AND all** or **OR all**.

Example:
To remove all spots that are not present in at least 85% of the spot maps, add the following filter criteria to the list:
Select the filter criteria **% of spot maps where protein is present**, choose the operator **>**, enter the value **85** and click **Add**.



Select filter criteria: % of spot maps where protein is present >
Value: 85

Filter Criteria		Value
% of spot maps where protein is present	>	85

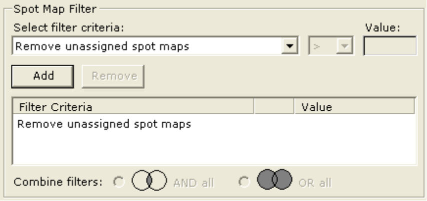
3 *Define Spot Map Filter criteria:*

a. Select filter criteria and **<**, **<=**, **=**, **>=** or **>** (only for **% of proteins present in spot map**) from the drop-down lists in the **Select filter criteria** field. For information on the different criteria, see section 5.4.3, Protein and spot map filter criteria.

b. Enter a value for the criteria in the **Value** field and click **Add** to add the filter criteria to the list. If required, repeat step 3a and 3b to add a new spot map filter criteria to the list.

c. Combine the filter criteria for the spot map filter by using the logical conditions **AND all** or **OR all**.

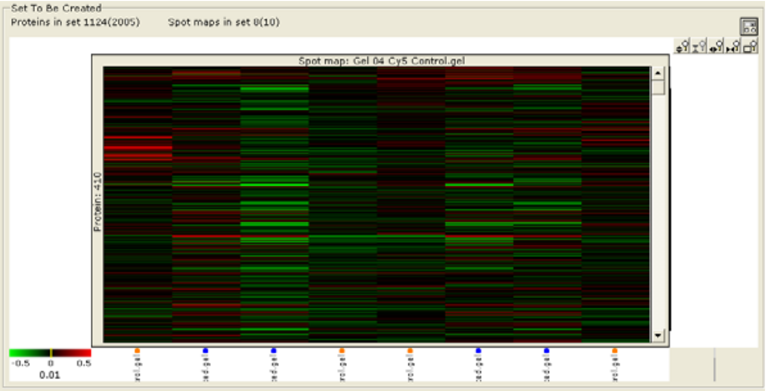
Example:
To remove all unassigned spot maps (spot maps contained within the Unassigned group in the experimental design), select the spot map filter **Remove unassigned samples** and click **Add** to add this to the filter criteria list.



Select filter criteria: Remove unassigned spot maps

Filter Criteria		Value
Remove unassigned spot maps		

4 Click **Apply Filter** to view the results of the filtering in the heat map below. The heat map will be updated showing only the proteins and spot maps that were included after the filter step. The **Proteins in set** and **Spot maps in set** fields will show how many proteins and spot maps that were included by the filter.



Set To Be Created

Proteins in set 1124(2005) Spot maps in set 8(10)

Spot map: Gel 04 Cy5 Control.gel

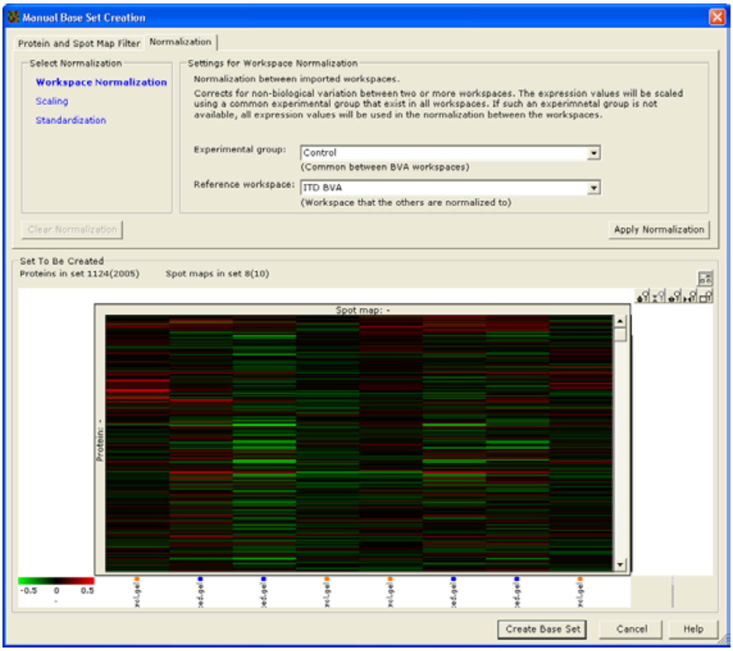
It is possible to edit the protein and spot map filter criteria and click **Apply Filter** again if you want to change any of the filter criteria. This procedure can be repeated until you are satisfied with the filters.

5 *Proceed with normalization of the data (step 6) only if* several BVA workspaces that do not use the same internal standard are included in the EDA workspace.

Otherwise, proceed with step 7.

6 *To normalize the data set:*

a. Click the **Normalization** tab to display the methods and options for normalization.



Normalization between imported workspaces.

Corrects for non-biological variation between two or more workspaces. The expression values will be scaled using a common experimental group that exist in all workspaces. If such an experimnetal group is not available, all expression values will be used in the normalization between the workspaces.

Experimental group: Control
(Common between BVA workspaces)

Reference workspace: ITD BVA
(Workspace that the others are normalized to)

Proteins in set 1124(2005)   Spot maps in set 8(10)

b. Select the appropriate normalization method to use in the Select Normalization area and enter settings in the Settings area.

See Appendix A, Normalization for information regarding normalization methods and settings.

c. Click **Apply Normalization**. The heat map will be updated showing only the proteins and spot maps that were included by the filter.

To clear any performed normalizations, click **Clear Normalization**.

7 Click **Create base set**. During the creation, a dialog showing the progress will be displayed.

8 When the base set has been created, the status **Base set created-Calculation is now possible** is displayed in the **Status** field of the **Base set creation** area in the **Setup** window. Proceed with section 5.5, Saving the EDA workspace.

5.4.3 Protein and spot map filter criteria

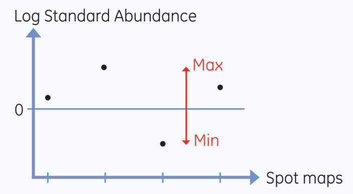
Protein Filter criteria

The available protein filter criteria when creating the base set is listed inTable 5-1.

Note: *When creating a set based on the results of calculations, the different calculation methods can also be used as protein filter criteria. If for example a Student's T-test was performed this will appear in the drop-down list for filter criteria and a value for filtering can be entered. See section 12.1.3 for all possible criteria.*

Table 5-1. Available protein filter criteria when creating the base set (general filter criteria).

Criteria	Value	Description
% of spot maps where protein is present	Numerical (%)	***Tip!*** *Use this criteria to remove proteins that have a lot of missing values among the spot maps.* Choose this criteria to include only those proteins that exist in a certain amount of spot maps in the data set. For example, if **>=** 80% is entered in the **Value** field, only proteins that have an expression value in >=80% of the spot maps (missing values are < 20%), will be included by the filter.
% of exp. groups where protein is present	Numerical (%)	***Tip!*** *Use this criteria to remove proteins that have a lot of missing values among the experimental groups.* Choose this criteria to include only those proteins that exist in a certain amount of experimental groups in the data set. For example, if >= 80% is entered in the **Value** field, only proteins that exist in >=80% of the experimental groups, will be included by the filter.
Standard deviation of log std. abundance	Numerical (range: 0-5)	Choose this criteria to include only those proteins with certain standard deviations. The standard deviation is a measure of the data spread and has the same unit as the observations (log standard abundance).

Criteria	Value	Description
Log std. abundance difference	Numerical (>0)	Choose this criteria to only include proteins with a certain log standard abundance difference (**Max**-**Min** difference), i.e. proteins that have large expression differences among the spot maps. 

Spot map filter criteria

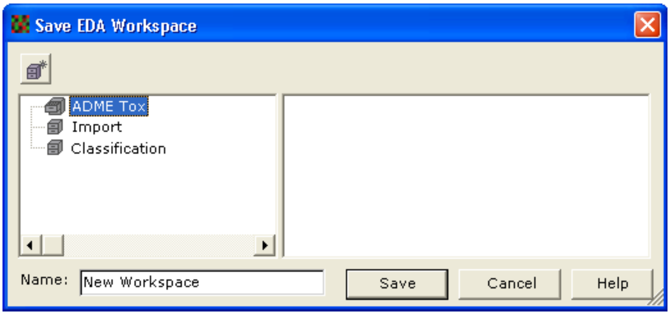
Criteria	Value	Description
% of proteins present in spot map	Numerical (%)	***Tip!*** *Use this criteria to remove spot maps that have a lot of missing protein expression values.* Choose this criteria to only include spot maps containing a certain amount of spots. For example, if >= 80% is entered in the **Value** field, only spot maps with at least 80% protein values (<20% missing values) are included by the filter.
Remove unassigned spot maps	n/a	Choose this criteria to remove all unassigned spot maps.

5.5 Saving the EDA workspace

When the base set has been created it is recommended to save the EDA workspace.

To save the EDA workspace:

1 Select **File:Save Workspace** in the menu bar. The **Save EDA workspace** dialog is displayed.



Save EDA Workspace

ADME Tox
Import
Classification

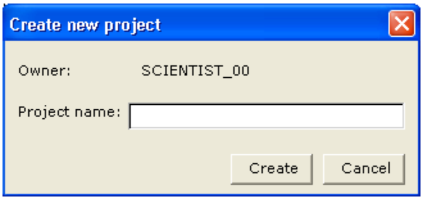
Name: New Workspace

2 Select the project in which to save the workspace.

Alternatively, create a new project in which to save your workspace as follows:

a. Click the **New project** icon.

b. The **Create new project** dialogis displayed. The **Owner** is the logged on user.



Create new project

Owner: SCIENTIST_00

Project name:

c. Enter a name for the project

d. Click **Create** to create the project and return to the **Save EDA workspace** dialog. The created project will be selected in the **Save EDA workspace** dialog.

3 Enter a name for the workspace in the **Name** field.

4 Click **Save**.

6 Calculations and Results - Overview

6.1 Overview

The second step in the workflow area, **Calculations**, is enabled when a base set has been created. This step includes setting up and performing calculations for a selected set of data.

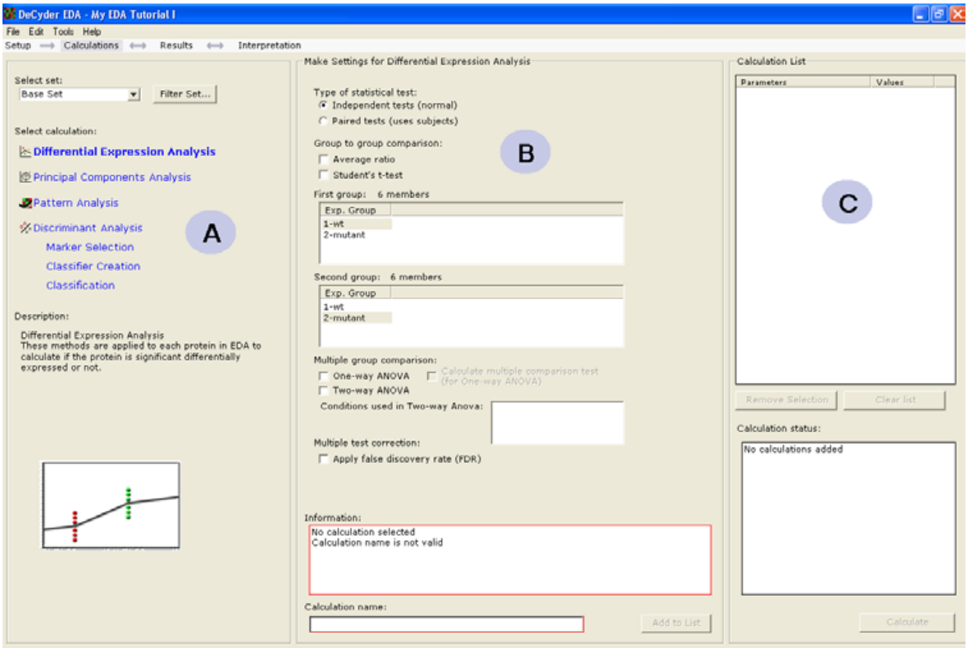
The workflow of the **Calculations** step is connected to the **Results** step. Usually, one or several calculations on the base set are set up and calculated in the **Calculations** step. The results of the calculations are then analyzed in the **Results** step and one or several new sets of data extracted from the analyses can be created.

It is then possible to:

- Return to the **Calculations** step and perform calculations on new or old sets with other settings or to perform new calculations
- Perform interpretation of the results

The calculations are set up, added to the calculation list and calculated in the ***Calculations*** *window, which consists of three main areas:*

- **Calculations (A)**
  Select a set on which to perform statistical analyses and select the type of statistical analysis to perform: **Differential Expression Analysis**, **Principal Component Analysis**, **Pattern Analysis** or **Discriminant Analysis**.
- **Make Settings (B)**
  Make settings for the analysis selected in the **Calculations** area and add the calculation to the calculation list. Add other statistical analyses to the calculation list by choosing a new analysis, entering settings and adding the calculation to the calculation list.
- **Calculation List (C)**
  Review the added calculations and perform the calculations in the list by clicking **Calculate**.



Select set: Base Set

Select calculation:

Description:

Differential Expression Analysis
These methods are applied to each protein in EDA to calculate if the protein is significant differentially expressed or not.

Make Settings for Differential Expression Analysis

Type of statistical test:
- ● Independent tests (normal)
- ○ Paired tests (uses subjects)

Group to group comparison:
- □ Average ratio
- □ Student's t-test

First group: 6 members
1-wt
2-mutant

Second group: 6 members
1-wt
2-mutant

Multiple group comparison:
- □ One-way ANOVA
- □ Calculate multiple comparison test (for One-way ANOVA)
- □ Two-way ANOVA

Conditions used in Two-way Anova:

Multiple test correction:
- □ Apply false discovery rate (FDR)

Information:
No calculation selected
Calculation name is not valid

Calculation name:

Calculation List

Calculation status:
No calculations added

Fig 6-1. Calculations window.

The statistical analyses are calculated in the order given in the calculation list. The status for a calculation is indicated in front of each calculation.

When the calculations have been performed, the results of the calculations are viewed and analyzed in the **Results** window.

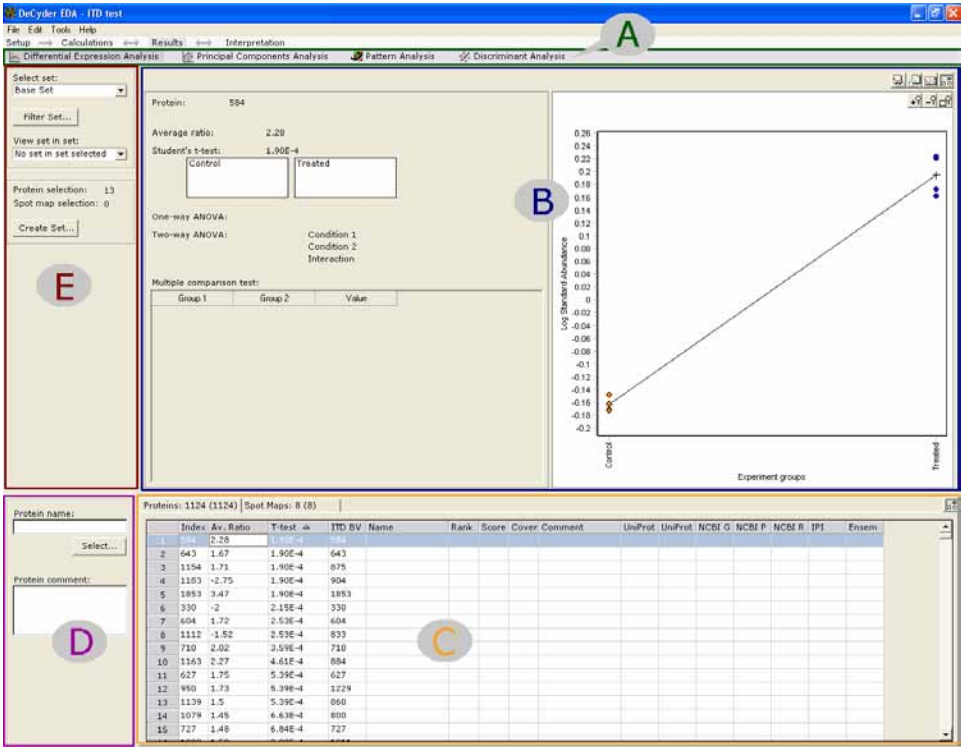
*There are four main areas in the **Results** window:*

- **Results bar** (A)
  Select which calculation results to display in the results view and protein/spot map table.

- **Results view (B)**
  View and analyze the selected calculation results.

- **Protein/Spot map table (C)** and **Protein/spot map details area (D)**
  The Protein and Spot map tables (C) show information on all proteins and spot maps in a table format. When highlighting a protein/spot map in the tables or in the results view, details on the selected protein/spot map will be displayed in the protein/spot map details areas (D).

- **Set area (E)**
  Select which set(s) to view in the results view and protein/spot map table and create new sets.



Select set: Base Set

View set in set: No set in set selected

Protein selection: 13
Spot map selection: 0

Protein: 584

Average ratio: 2.28
Student's t-test: 1.90E-4

Control | Treated

One-way ANOVA:

Two-way ANOVA: Condition 1
Condition 2
Interaction

Multiple comparison test:

Group 1	Group 2	Value

Protein name:

Protein comment:

Proteins: 1124 (1124) | Spot Maps: 8 (8)

	Index	Av. Ratio	T-test	ITD BV	Name	Rank	Score	Cover	Comment	UniProt	UniProt	NCBI G	NCBI P	NCBI R	IPI	Ensem
1	584	2.28	1.90E-4	584												
2	643	1.67	1.90E-4	643												
3	1154	1.71	1.90E-4	875												
4	1103	-2.75	1.90E-4	904												
5	1853	3.47	1.90E-4	1853												
6	330	-2	2.15E-4	330												
7	604	1.72	2.53E-4	604												
8	1112	-1.52	2.53E-4	833												
9	710	2.02	3.59E-4	710												
10	1163	2.27	4.61E-4	884												
11	627	1.75	5.39E-4	627												
12	950	1.73	5.39E-4	1229												
13	1139	1.5	5.39E-4	860												
14	1079	1.45	6.63E-4	800												
15	727	1.48	6.84E-4	727												

Fig 6-2. Results window.

New sets can be created by selecting data of interest or by filtering the data set. Created sets can also be combined into new sets to, for example, extract a subset of proteins and spot maps from the sets.

It is then possible to go back to the **Calculations** step and perform more calculations or go to the **Interpretation** step and perform biological interpretation.

For example workflows, work through the tutorials in Chapter 15 and 16.

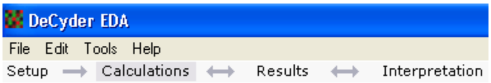
See section 6.2 for a more detailed workflow in the **Calculations** and **Results** steps.

6.2 Workflow for Calculations and Results

The workflow for the **Calculations** and **Results** steps are outlined in sections 6.2.1-6.2.6.

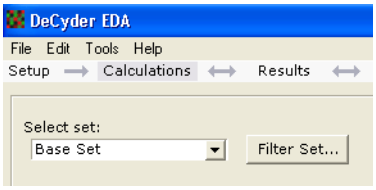
6.2.1 Select the set on which to perform calculations

1 If not already in the **Calculations** window, click **Calculations** in the workflow area.



The **Calculations** window opens.

2 In the **Select set** drop-down list, select the **Set** on which to perform statistical analyses.

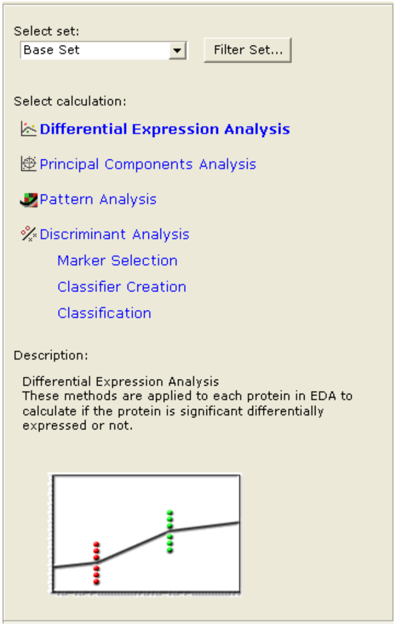


Select set: Base Set

Note: *When performing calculations for the first time only the created base set is available.*

6.2.2 Select calculation method

1 Click the appropriate method name (displayed in blue) in the left panel.



Differential Expression Analysis
These methods are applied to each protein in EDA to calculate if the protein is significant differentially expressed or not.

2 The settings for the selected method are displayed in the middle panel.

6.2.3 Make settings and add calculations to the Calculation List

Depending on which main analysis was selected in section 6.2.2, different analyses can be set up and settings entered or changed for the selected analyses.

To select settings:

1 Enter settings for the selected calculation in the middle panel (**Make settings** area) of the **Calculations** window.

For information on how to enter settings for the different analyses available in EDA, see Chapter 7-10.

For detailed information about the different methods and more advanced settings for the different methods, refer to Appendix B, Statistics and algorithms - Introduction and the Online help.

2 Add the calculation to the calculation list by entering a name for the calculation in the **Calculation** name field and clicking **Add to List**.

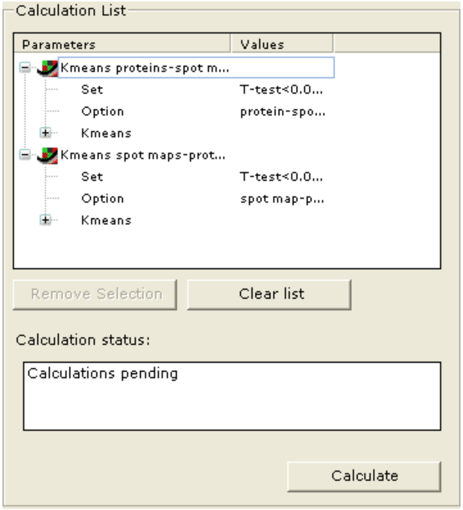
3 If required, repeat the steps in sections 6.2.2 and 6.2.3 to add more calculations to the calculation list. The calculations can be performed in any specific order. The calculations are performed sequentially one at a time when clicking **Calculate**.

Note: *Only one differential expression calculation/set and one set of hierarchical clustering calculations/pattern to be calculated per set can be added at a time. If adding further calculations, the previous ones will be overwritten. For all other analyses, it is possible to add several calculations/set with different settings.*

6.2.4 Perform the calculations in the Calculation List

When all calculations have been added to the calculation list:

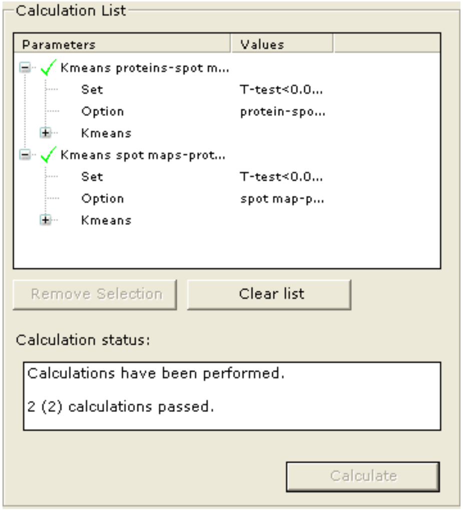
1 In the **Calculation List** area, click **Calculate** to start the calculation.



Parameters	Values
Kmeans proteins-spot m...	
Set	T-test<0.0...
Option	protein-spo...
Kmeans	
Kmeans spot maps-prot...	
Set	T-test<0.0...
Option	spot map-p...
Kmeans	

Calculations pending

During calculation the status of each calculation is indicated by an icon in front of the calculation. The status of the calculations is also displayed in the **Calculation status** field.



Parameters	Values
✓ Kmeans proteins-spot m...	
Set	T-test<0.0...
Option	protein-spo...
Kmeans	
✓ Kmeans spot maps-prot...	
Set	T-test<0.0...
Option	spot map-p...
Kmeans	

Calculations have been performed.

2 (2) calculations passed.

2 Calculation status icons indicate when the calculation is in progress, has finished successfully, has failed or has been canceled. The following status

icons may appear in front of the calculation:

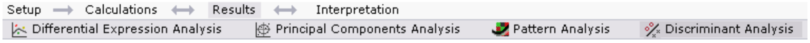
Icon	Description
●	The calculation is in progress.
✓	The calculation has successfully finished.
✗	The calculation has been cancelled.
●	The calculation has failed.

Select **Results** in the workflow area to view the results of the calculations. For information on how to analyze the results for the different analyses, see Chapter 7-10.

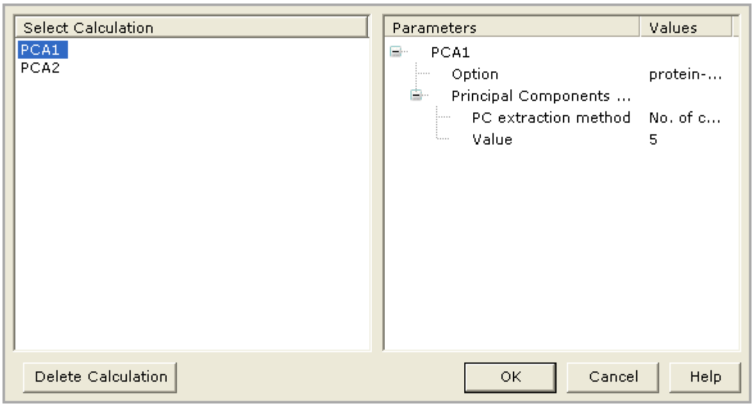
Note: *It is always possible to return to the **Calculations** step from the **Results** step in order to perform additional calculations.*

6.2.5 Display the results of an analysis

Click the appropriate calculation in the results bar to view the results of the corresponding calculations (if performed) in the results view and the protein/spot map table. Depending on which analysis is selected, the results view and protein/spot map table will appear different. The protein details area and set area are common to all the analyses.



If Principal Component Analysis, Pattern Analysis (partitioning clustering) or Discriminant Analysis (Feature selection or Classifier creation) were selected, the appropriate calculation result must be selected in the **Calculation result** field, if several calculations have been performed.



Select Calculation
PCA1
PCA2

Parameters	Values
PCA1	
Option	protein-...
Principal Components ...	
PC extraction method	No. of c...
Value	5

6.2.6 Analyze the results

1 Analyze the results for the selected calculation in the results bar.

For information on how to analyze the results of the different calculations, see Chapter 7-10.

2 Create new sets by selecting data from the protein/spot map table and/or results view or by filtering of the results.

3 Return to the **Calculations** step and perform calculations on new or old sets with other settings or perform new calculations (repeat section 6.2).

or

Go to the **Interpretation** step and perform biological interpretation on the set with the proteins of interest (see Chapter 11 for information on interpretation).

6.3 Calculations available in EDA

The different calculations (statistical analyses) that can be performed in EDA are divided into 4 main groups: **Differential Expression Analysis**, **Principal Component Analysis**, **Pattern Analysis** and **Discriminant analysis**, each containing a number of sub-analyses. Table 6-1 summarizes the 4 main groups of analyses available in EDA.

Main analysis	Example of biological queries	Information on calculation settings and result analysis
Differential Expression Analysis	Investigate differential expression between two or more experimental groups. The tests can be independent or paired.	See Chapter 7, Calculation and Results - Differential Expression Analysis.
Principal Component Analysis	Identify outliers and initial groupings of data.	See Chapter 8, Calculation and Results - Principal Component Analysis.
Pattern Analysis	Investigate if any patterns exist among proteins or spot maps.	See Chapter 9, Calculation and Results - Pattern Analysis.
Discriminant analysis	Identify diagnostic or prognostic markers. Create classifier and classify samples into known class.	See Chapter 10, Calculation and Results - Discriminant Analysis.

Table 6-1. Summary of the available analyses in EDA.

The analyses can be performed in any order. To gain an understanding of how and when the different calculation methods can be used, work through the tutorials in Chapter 15 and 16.

For detailed information about the different methods, see Appendix B or the Online help.

7 Calculation and Results - Differential Expression Analysis

7.1 Introduction

This chapter gives an overview of how to:

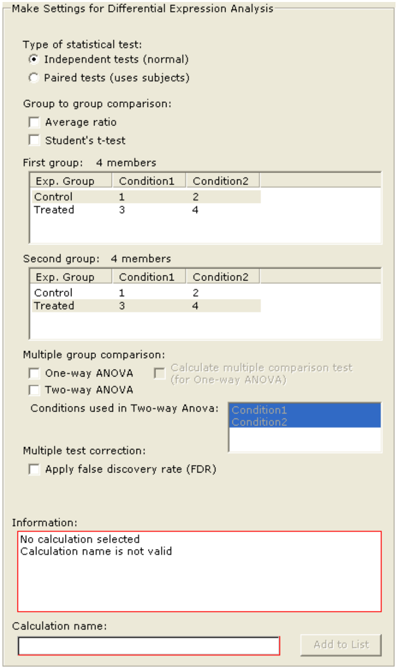
- Make settings for differential expression analysis in the **Make settings for differential analysis** area of the **Calculations** window
- Analyze the results for the differential expression analysis

7.2 Make settings for differential expression analysis

7.2.1 Overview

Usually, the differential expression analysis is performed first in order to find significantly differentially expressed proteins.

The settings for differential expression analysis contain different sub-analyses and settings. Depending on the experimental setup, Average ratio, Student's T-test or ANOVA analyses can be selected.



Make Settings for Differential Expression Analysis

Type of statistical test:
● Independent tests (normal)
○ Paired tests (uses subjects)

Group to group comparison:
□ Average ratio
□ Student's t-test

First group: 4 members

Exp. Group	Condition1	Condition2
Control	1	2
Treated	3	4

Second group: 4 members

Exp. Group	Condition1	Condition2
Control	1	2
Treated	3	4

Multiple group comparison:
□ One-way ANOVA □ Calculate multiple comparison test (for One-way ANOVA)
□ Two-way ANOVA

Conditions used in Two-way Anova: Condition1 Condition2

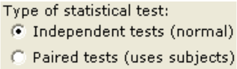
Multiple test correction:
□ Apply false discovery rate (FDR)

Information:
No calculation selected
Calculation name is not valid

Calculation name:

Note: *For detailed information on analyses and settings, see Appendix C, Statistics and algorithms - Differential Expression Analysis or the online help. The guidelines below give only a brief overview.*

7.2.2 Make settings for Differential Expression Analysis



Type of statistical test:

● Independent tests (normal)

○ Paired tests (uses subjects)

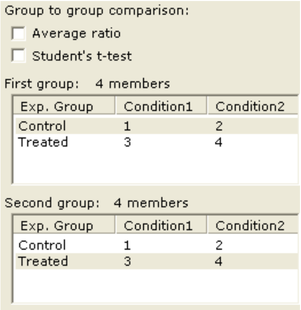
1 Select the **Type of statistical test** to perform by choosing the appropriate radio button in the **Type of statistical test** area.

Type of test	Guidelines
Independent test (normal)	Independent tests are general techniques that can be used to test whether standardized protein abundance differs between groups and does not require the groups to be paired in any way, or even to be of equal sizes.
Paired test (uses subjects)	Paired test can be used when each data point in one group corresponds to a matching data point in the other group(s). A typical example would be the same group of patients before and after a treatment.

2 *If the protein expression of exactly two groups or two population of groups are to be compared:*

Select which analyses to perform by checking the appropriate boxes in the **Group to group comparison** area.

If performing a **Student's T-test**, select which groups are to be compared by clicking a group/population of groups in the **First group** area and another group/population of groups in the **Second group** area.



Group to group comparison:

□ Average ratio

□ Student's t-test

First group: 4 members

Exp. Group	Condition1	Condition2
Control	1	2
Treated	3	4

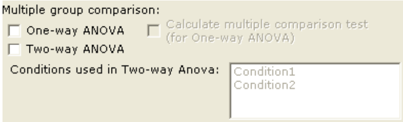
Second group: 4 members

Exp. Group	Condition1	Condition2
Control	1	2
Treated	3	4

Type of analysis	Brief description
Average ratio	Calculates the difference in the standardized abundance between 2 protein spot groups.
Student's T-test	Student's T-test is used to test the hypothesis that a variable differs between two groups. There must be at least two members in each group. Otherwise, only the average ratio can be calculated.

3 *If more than two groups are available and the protein expression among all groups is to be compared:*

Select which analyses to perform by checking the appropriate boxes in the **Multiple group comparison** area.



Multiple group comparison:

□ One-way ANOVA □ Calculate multiple comparison test (for One-way ANOVA)

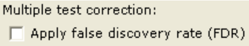
□ Two-way ANOVA

Conditions used in Two-way Anova: Condition1 Condition2

If performing **Two-way ANOVA** analysis, select the two conditions to use in the **Conditions used in two-way anova** field.

Type of analysis	Brief description
One-Way ANOVA	One-way ANOVA is used to test for differences in standardized abundance among all groups. The test will not indicate which groups are different from which other groups, just that there is an overall difference. All groups that have at least two members will be included in the calculation.
Calculate multiple comparison test (for One-way ANOVA)	Check this box to get an indication of which groups are different in the One-Way ANOVA result presentation.
Two-way ANOVA	Two-Way ANOVA calculates the significance of the difference between groups with the same condition 2 and different condition 1 values (Two–Way ANOVA Condition 1) and the other way around (Two–Way ANOVA Condition 2). The Two-Way ANOVA analysis also calculates the significance value of the mutual effect of the two factors (Two–Way ANOVA Interaction).

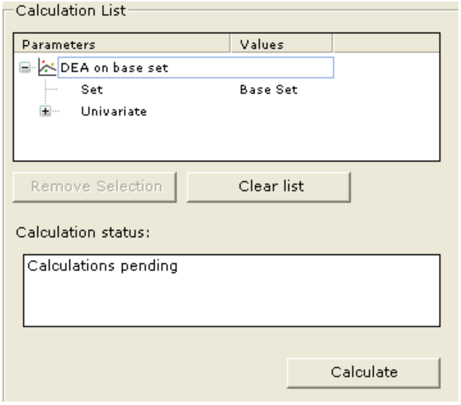
4 Check the **Apply false discovery rate (FDR)** box in the **Multiple test correction** area if the Student's T-test or ANOVA values for each protein should be adjusted to keep the overall error rate as low as possible.



Multiple test correction:

□ Apply false discovery rate (FDR)

5 Enter a name for the calculation in the **Calculation name** field and click **Add to List**. The calculation is added to the calculation list.



6 Click **Calculate** to perform the calculation

or

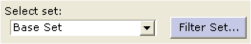
Add other types of calculations (PCA, Pattern Analysis and Discriminant Analysis) to the **Calculation List** (see Chapter 6 for information on the workflow).

Note: *Only one differential expression analysis calculation/set can be added to the list. If adding another one to the same set, a dialog will appear, asking if you want to overwrite the previous analysis.*

7 When a calculation has finished, this will be indicated by a status icon in front of the calculation and displayed in the **Calculation status** field. The following status icons may appear in front of the calculations:

Icon	Description
🟢	The calculation is in progress.
✓	The calculation has successfully finished.
✗	The calculation has been cancelled.
🔴	The calculation has failed.

For information on how to analyze the results, see section 7.3.

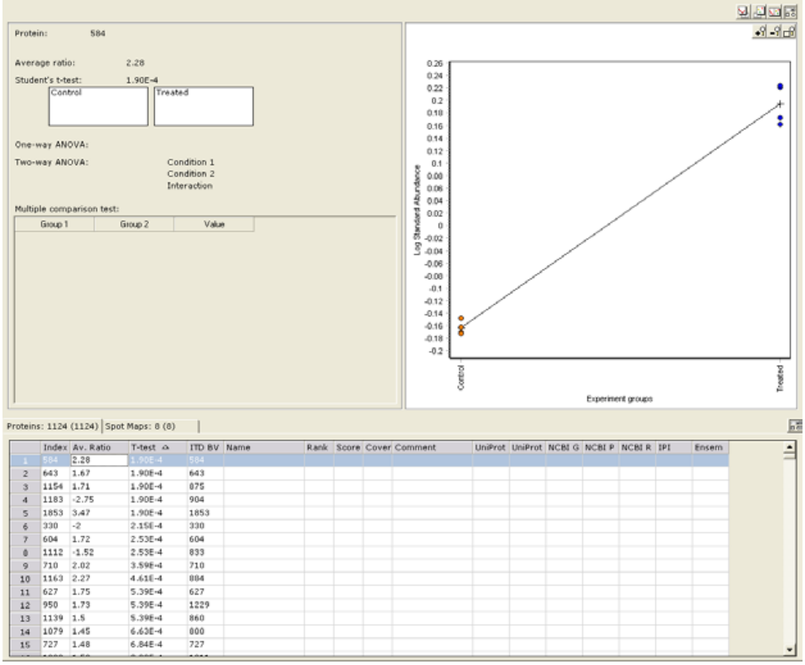

Select set: Base Set

Note: *If the results are to be filtered using one or several of the statistical analysis results, it is possible to do this in the Calculations window as well as in the Results window by clicking* ***Filter set...*** *in the Set area. For information on how to perform filtering, see section 7.3.1.*

7.3 Analyze the results of the differential expression analysis

The results from the **Differential Expression Analysis** are displayed in the protein table and in the results view. The protein/spot map table summarizes the results for all proteins (**Proteins** tab, protein table) and spot maps (**Spot Maps** tab, spot map table). To view detailed results for a protein, select the appropriate protein in the protein table. The detailed results will be shown in the results view.

Tip! *For detailed information about the settings in the graph view, click in the graph view and press F1 to open the online help for this area.*



Protein: 584

Average ratio: 2.28

Student's t-test: 1.90E-4

Control | Treated

One-way ANOVA:

Two-way ANOVA: Condition 1
Condition 2
Interaction

Multiple comparison test:

Group 1	Group 2	Value

Proteins: 1124 (1124) | Spot Maps: 8 (8)

	Index	Av. Ratio	T-test	ITD BV	Name	Rank	Score	Cover	Comment	UniProt	UniProt	NCBI G	NCBI P	NCBI R	IPI	Ensem
1	584	2.28	1.90E-4	584												
2	643	1.67	1.90E-4	643												
3	1154	1.71	1.90E-4	875												
4	1183	-2.75	1.90E-4	904												
5	1853	3.47	1.90E-4	1853												
6	330	-2	2.15E-4	330												
7	604	1.72	2.53E-4	604												
8	1112	-1.52	2.53E-4	833												
9	710	2.02	3.59E-4	710												
10	1163	2.27	4.61E-4	884												
11	627	1.75	5.39E-4	627												
12	950	1.73	5.39E-4	1229												
13	1139	1.5	5.39E-4	860												
14	1079	1.45	6.63E-4	800												
15	727	1.48	6.84E-4	727												

The analysis of the differential expression analysis results are performed by extracting significantly differentially expressed proteins and/or proteins with a certain fold change and creating a new set containing these proteins. This can be done by:

- *Filtering the results (normally performed)*
  Filter the results with respect to p-value and create a new set including the filtered proteins (see section 7.3.1)

  or

- *Sorting the results and manually selecting proteins*
  Sort the results based on p-value, manually view and select interesting proteins in the protein table and create a set with the selected proteins (see section 7.3.2)

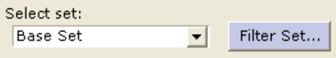
7.3.1 Filter the results

Filter the results if you know that you want to extract all proteins with certain p-values and create a new set containing only these proteins. Any of the differential expression analysis calculations performed can be used in the filtering process.

Note: *Filtering of the results can be performed in both the Results window and Calculations window.*

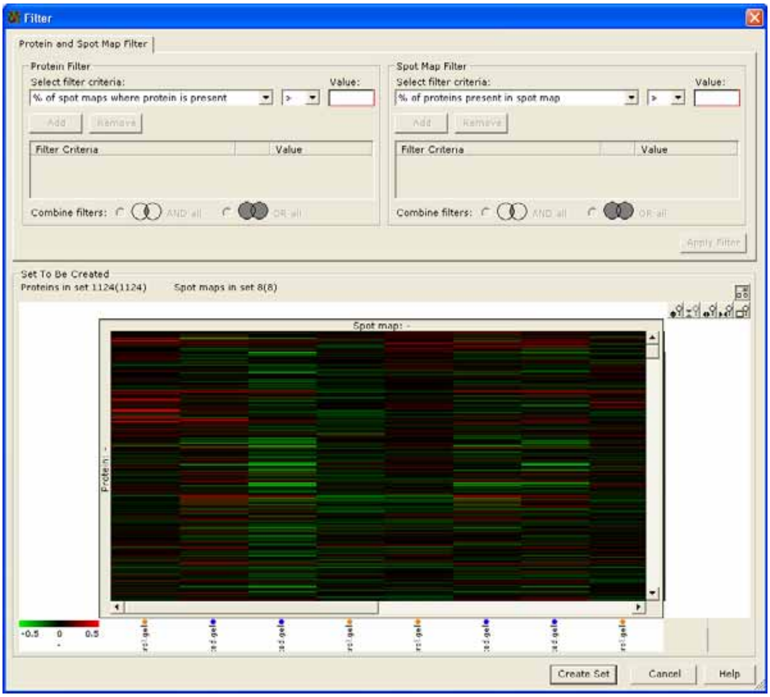
To filter the results:

1 Click **Filter Set...** in the **Set area** of the **Results** window or **Calculations** window**.**



Select set: Base Set

The **Filter** dialog opens.



Filter

Protein and Spot Map Filter

Protein Filter
Select filter criteria: % of spot maps where protein is present >
Value:

Spot Map Filter
Select filter criteria: % of proteins present in spot map >
Value:

Set To Be Created
Proteins in set 1124(1124) Spot maps in set 8(8)

2 In the **Protein filter** area:

a. Select the filter criteria to base the filtering on from the first drop-down list in the **Select filter criteria** field. The filter to choose should be one or several of the performed differential expression analysis calculations. Table 7-1 lists the possible differential expression analysis filter criteria. The calculation must have been performed in order to appear in the list.

Note: *It is also possible to filter data using general filter criteria. To view a description of the general protein filter criteria, see section 5.4.3.*

Criteria	Description
Average ratio **Paired Average Ratio**	Use this criteria to extract proteins with certain fold changes. If a paired test was performed, the Paired Average Ratio will appear.
Student's T-test **Paired Student's T-test**	Use this criteria to extract proteins with certain Student's T-test p-values (if an independent test was performed). If a paired test was performed, the **Paired Student's T-test** will appear.
One-Way ANOVA **RM One-Way ANOVA**	Use this criteria to extract proteins with certain **One-Way ANOVA** p-values (if an independent test was performed). If a paired test was performed, the **RM One-Way ANOVA** will appear.
Two-Way ANOVA, condition 1 **Two-Way ANOVA, condition 2** **Two-Way ANOVA, condition interaction**	Use this criteria to extract proteins with certain **Two-Way ANOVA** p-values (if an independent test was performed). The **Two-Way ANOVA** calculation gives three p-values: **Two-Way ANOVA, condition 1**, **Two-Way ANOVA, condition 2** and **Two-Way ANOVA, condition interaction**. All these criteria will be available and can be used for filtering.
RM Two-Way ANOVA, condition 1 **RM Two-Way ANOVA, condition 2** **RM Two-Way ANOVA, condition interaction**	Use this criteria to extract proteins with certain **RM Two-Way ANOVA** p-values (if an independent test was performed). The **RM Two-Way ANOVA** calculation gives three p-values: **RM Two-Way ANOVA, condition 1**, **RM Two-Way ANOVA, condition 2** and **RM Two-Way ANOVA, condition interaction**. All these criteria will be available and can be used for filtering.

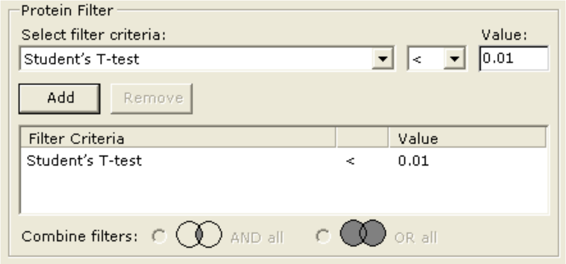
Table 7-1. Differential expression analysis filter criteria.

b. Select one of the operators **<**, **<=**, **=**, **>=** or **>** from the second drop-down list in the **Select filter criteria** field.

c. Enter a value for the criteria.

For example, the Student's T-test and a value of 0.01 can be used in combination with the operator **<** to extract only those proteins with a p-value less than 0.01.

3 Click **Add** to add the filter criteria to the list.


Protein Filter

Select filter criteria: Student's T-test < Value: 0.01

Filter Criteria		Value
Student's T-test	<	0.01

Combine filters: ○ AND all ○ OR all

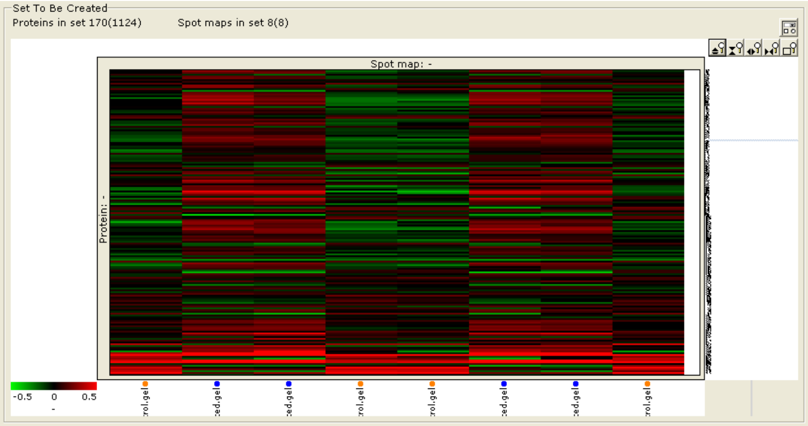
4 *If appropriate*, repeat step 2-3 and select another filter to add to the filter criteria list (for example Average ratio > 2, to extract proteins with a greater than 2-fold change in expression).

5 *If more than one filter was added to the filter criteria list:*
Combine the filters by selecting either the **AND all** or **OR all** radio button.

Radio button	Description
AND all	Includes only those proteins extracted by all filter criteria.
OR all	Includes those proteins extracted by at least one of the filter criteria.

6 Click **Apply Filter** to apply the filter to the set.

7 The number of proteins and spot maps extracted by the filter are displayed in the **Set To Be Created** area together with the heat map. For information on how to zoom in and how to change settings in the heat map, see section 3.3, The heat map.



Set To Be Created

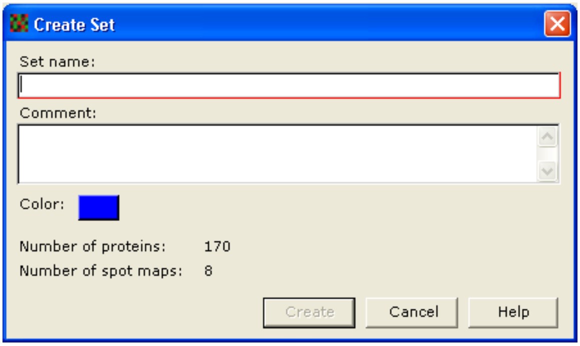
Proteins in set 170(1124)   Spot maps in set 8(8)

8 If you are satisfied with the filter, proceed with the next step. Otherwise, edit the filter by selecting the filter criteria, clicking **Remove filter** and adding a new filter by repeating steps 2, 3 and 6.

For example, if the filter extracted too many proteins, add a filter that extracts proteins with even lower p-values.

9 Click **Create set**.

The **Create Set** dialog opens showing the number of proteins and spot maps selected by the filter.



Create Set

Set name:

Comment:

Color: 🟦

Number of proteins: 170

Number of spot maps: 8

10 Enter a name for the set and, if required, a comment.

11 If desired, change the color for the set by clicking the color button and choosing the appropriate color.

Tip! *Different colors for the sets facilitates the interpretation of the results of different analyses in the results step.*

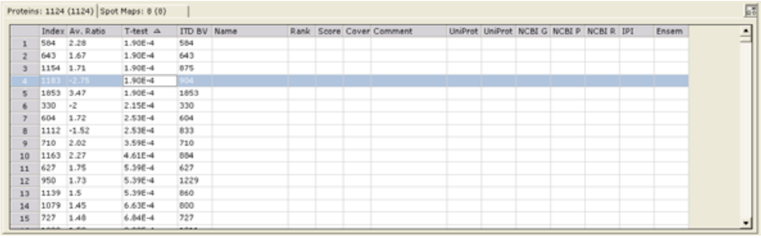
12 Click **Create** to create the set. Create more sets, perform more calculations or go to the **Interpretation** step.

7.3.2 Sort the results and manually select proteins

This can be performed to manually select proteins of interest. The results of the performed differential expression analysis calculations (Average ratio, Student's T-test or the ANOVA-analyses) are displayed in the corresponding column in the Protein table.

To sort the results, select interesting proteins and create a new set:

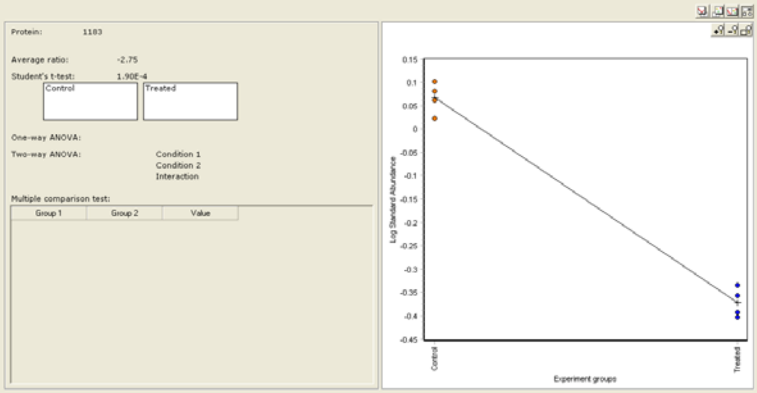
1 To sort the proteins based on a certain analysis, click the appropriate column header. The table is sorted according to the column. An arrow is displayed in the column header indicating that the table has been sorted according to that column. Click the column header again to reverse the sorting order.



Proteins: 1124 (1124) | Spot Maps: 8 (8)

	Index	Av. Ratio	T-test ▲	ITD BV	Name	Rank	Score	Cover	Comment	UniProt	UniProt	NCBI G	NCBI P	NCBI R	IPI	Ensem
1	584	2.28	1.90E-4	584												
2	643	1.67	1.90E-4	643												
3	1154	1.71	1.90E-4	875												
4	1183	-2.75	1.90E-4	904												
5	1853	3.47	1.90E-4	1853												
6	330	-2	2.15E-4	330												
7	604	1.72	2.53E-4	604												
8	1112	-1.52	2.53E-4	833												
9	710	2.02	3.59E-4	710												
10	1163	2.27	4.61E-4	884												
11	627	1.75	5.39E-4	627												
12	950	1.73	5.39E-4	1229												
13	1139	1.5	5.39E-4	860												
14	1079	1.45	6.63E-4	800												
15	727	1.48	6.84E-4	727												

2 To view detailed results for a protein, select it in the table. The results of the analyses are displayed in the results view.



Protein: 1183

Average ratio: -2.75

Student's t-test: 1.90E-4

Control

Treated

One-way ANOVA:

Two-way ANOVA: Condition 1
Condition 2
Interaction

Multiple comparison test:

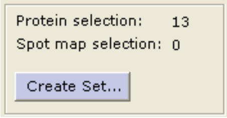
Group 1	Group 2	Value

The left panel shows the results in a numerical form for all analyses performed on the protein. The right panel shows a graph of the expression profile over the experimental groups by default. The settings for the graph can be altered.

For more information on graph settings and zooming, press **F1** when the results view is in focus to open the online help for the results view.

3 View detailed results for other proteins by selecting these proteins in the protein table.

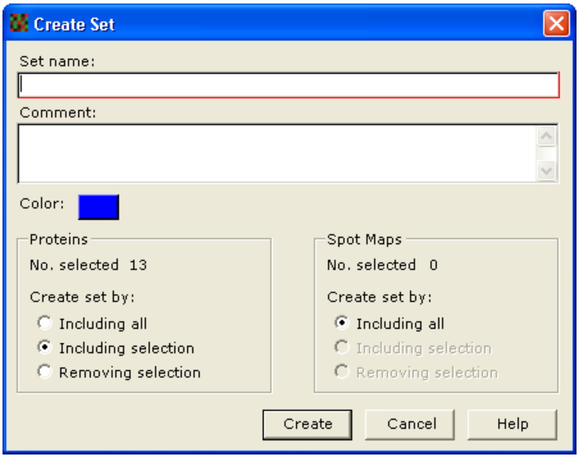
4 When proteins of interest have been found, select the proteins in the table using the **Ctrl** and **Shift** buttons. The number of selected proteins and spot maps are shown in the set area.



Protein selection: 13
Spot map selection: 0

5 Click **Create Set...** in the set area of the **Results** window to create a new set.

6 The **Create Set** dialog opens showing the number of proteins and spot maps selected.



Create Set

Set name:

Comment:

Color: 🟦

Proteins
No. selected 13
Create set by:
○ Including all
● Including selection
○ Removing selection

Spot Maps
No. selected 0
Create set by:
● Including all
○ Including selection
○ Removing selection

7 Enter a name for the set and, if required, a comment.

8 If required, change the color for the set by clicking the colored button and choosing the appropriate color.

Tip! *Different colors for the sets facilitates the interpretation of the results of different analyses in the results step.*

9 In the **Proteins** area, make sure that the **Including selection** radio button is selected.

In the **Spot Maps** area, select the **Including all** radio button.

10 Click **Create** to create the set. Create more sets, perform more calculations or go to the **Interpretation** step.

8 Calculation and Results - Principal Component Analysis

8.1 Introduction

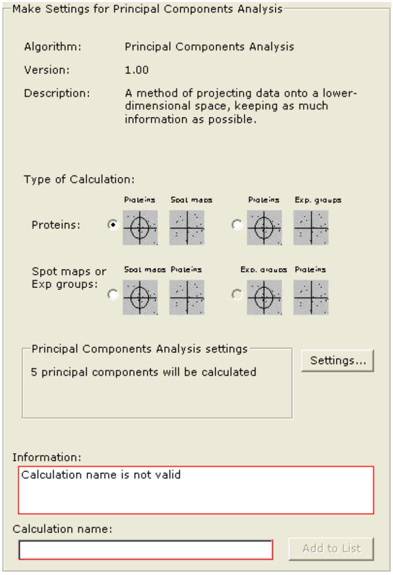
This chapter gives an overview of how to:

- Make settings for the different analyses in the **Make settings for Principal Component Analysis** area of the **Calculations** window
- Analyze the results of the Principal Component Analysis (PCA) calculations in the **Results** window

8.2 Make settings for PCA

The settings for PCA includes selecting what type of overview to produce and, if required, changing the settings for the PCA algorithm. This analysis is usually performed on the set with significantly differentially expressed proteins (created when the results of the differential expression analysis were analyzed) but can be performed on any set.

Note: *If the set on which PCA is performed contains too many missing values, the calculation will fail. This will be indicated by an icon in front of the calculation and a message with why the calculation failed will be displayed in the Calculation Status area. If your set contains many missing values, a new set where the missing values have been removed should be created before performing PCA. See section 12.1.2, Create a set by filtering data.*



Make Settings for Principal Components Analysis

Algorithm: Principal Components Analysis

Version: 1.00

Description: A method of projecting data onto a lower-dimensional space, keeping as much information as possible.

Type of Calculation:

Proteins: ● ○

Spot maps or Exp groups: ○ ○

Principal Components Analysis settings

5 principal components will be calculated

Information:

Calculation name is not valid

Calculation name:

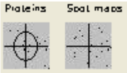
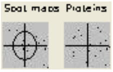
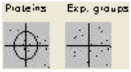
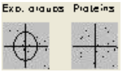
To make settings for PCA:

1 Select the type of calculation to perform by choosing the appropriate radio button in the **Type of calculation** area.

The icons to the right of each radio button show the type of overview that will be produced. The table below lists brief descriptions of the different overviews.

Note: *For each option, a separate calculation must be added to the calculation list (e.g. set up one calculation that calculates the overview of proteins-spot maps and set up another calculation that calculates the overview of spot maps-proteins).*

See Appendix D for detailed information about PCA.

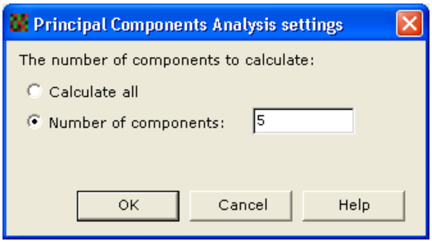
Type of analysis	**Brief description**
Proteins - Spot maps:* 	Use this analysis to find protein outliers and perform a ***rough*** comparison of the relationship between proteins and spot maps.
Spot maps - Proteins:* 	Use this analysis to check if there are any spot map outliers. Replica spot maps (spot maps in the same experimental group) should be grouped together in the left plot when viewing the results. It is also possible to perform a ***rough*** comparison of the relationship between spot maps and proteins.
Proteins - Exp. groups:** 	Use this analysis to find protein outliers and to perform a ***rough*** comparison of the relationship between proteins and experimental groups.
Exp. groups - Proteins:** 	Use this analysis to view the grouping of the experimental groups and to perform a ***rough*** comparison of the relationship between experimental groups and proteins.

* It is recommended to start with these analyses to produce an initial overview of the proteins and spot maps in the data set.

** The protein expression for an experimental group is calculated as the mean of the protein's expression on the spot maps in the experimental group. Therefore, it is recommended to check that no spot map outliers exist in the different experimental groups by performing PCA on spot maps - proteins before performing this analysis.

2 In the **Principal Component Analysis settings** area, the default settings for the analysis are displayed. These settings can normally be used.

However, if you want to change the PCA algorithm settings, click **Settings** to open the **Principal Component Analysis settings** dialog. See Appendix D for information on PCA and settings.



Principal Components Analysis settings

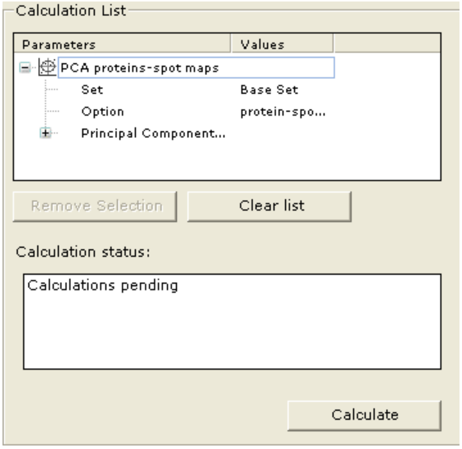
The number of components to calculate:

○ Calculate all

● Number of components: 5

3 Enter a name for the calculation in the **Calculation name** field and click **Add to List**.

The calculation is added to the calculation list.



Calculation List

Parameters	Values
PCA proteins-spot maps	
Set	Base Set
Option	protein-spo...
Principal Component...	

Calculation status:

Calculations pending

4 Click **Calculate** to perform the calculation.

or

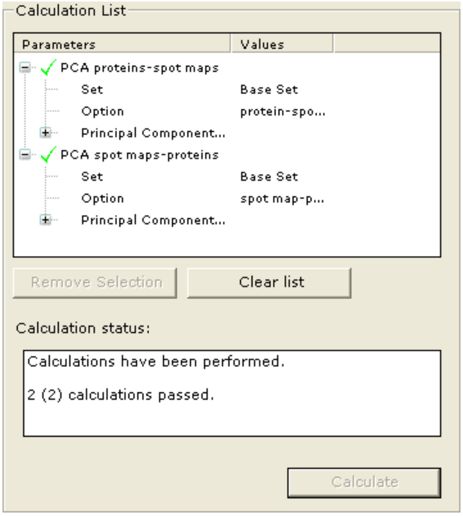
Add more PCA calculations or other types of calculations to the **Calculation** list (see Chapter 6 for information about the workflow).

Note: *It is recommended to add a calculation on proteins versus spot maps and a calculation on spot maps versus proteins to check that no protein mis-matches or spot map outliers exist.*

5 When a calculation has finished, this will be indicated by a status icon in front of the calculation. The following status icons may appear in front of the calculations:

Icon	Description
🟢	The calculation is in progress.
✓	The calculation has successfully finished.
✗	The calculation has been cancelled.
🔴	The calculation has failed.

The status of the calculations will also be displayed in the **Calculation status** field.



Parameters	Values
✓ PCA proteins-spot maps	
Set	Base Set
Option	protein-spo...
Principal Component...	
✓ PCA spot maps-proteins	
Set	Base Set
Option	spot map-p...
Principal Component...	

Calculation status:

Calculations have been performed.

2 (2) calculations passed.

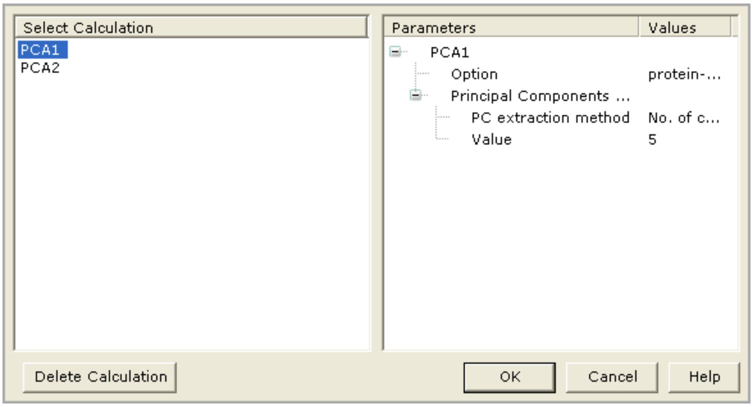
For information about how to analyze the results, see section 8.3.

8.3 Analyze the results of the PCA calculation

8.3.1 Select the calculation from which to view the results

1 Select the **Results** step in the workflow area and then **Principal Component Analysis** in the Results bar. The results for the PCA calculation in the **Calculation result** field are displayed in the results view.

2 If you want to view the results for another PCA calculation (if several were performed), click the **Calculation result** arrow button to display the **Select calculation** pop-up dialog.



Select Calculation
PCA1
PCA2

Parameters	Values
PCA1	
Option	protein-...
Principal Components ...	
PC extraction method	No. of c...
Value	5

3 Select the calculation for which results to display in the **Select calculation** column. The values for the parameters in the calculation are shown in the right panel (**Parameters** and **Values** columns).

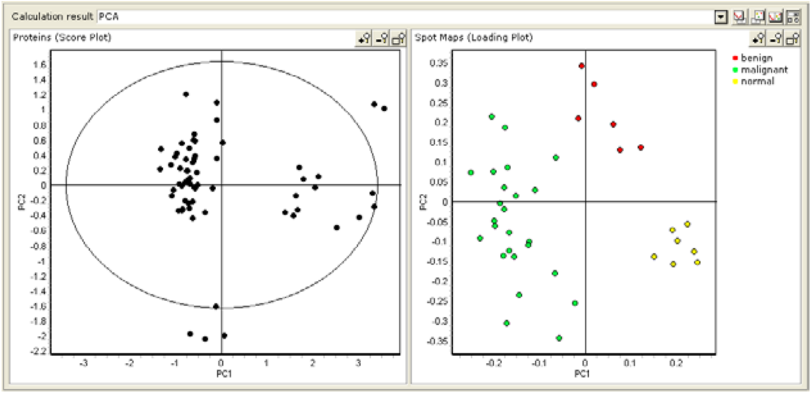
Tip! *If your calculation does not appear in the list, make sure the correct set is selected in the Select set field.*

4 Click **OK** to display the results for the selected calculation in the results view.

8.3.2 Overview of PCA results

Usually, the PCA has been performed on a smaller set than the base set, containing proteins extracted in the differential expression analysis. However, a PCA of ***proteins versus spot maps*** can be performed at the beginning to get an initial overview of the data set.

The results of the selected calculation in the **Calculation result** field are displayed in the form of a score plot and a loading plot.



Calculation result PCA

Fig 8-1. The results of a PCA of proteins versus spot maps.

Depending on the settings in the performed calculation, either proteins or spot maps/experimental groups are shown in the score plot (left plot) and spot maps/experimental groups or proteins are shown in the loading plot (right plot).

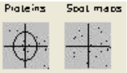
By default, the plots display the results in 2D-space (principal component 1, PC1, and principal component 2, PC2 on the axes). If different colors have been assigned to the experimental groups, it is possible to view which spot maps belong to which experimental groups in the plot with spot maps (see Fig. 8-1).

The results will indicate if there are any outliers in the data and also the relationship between proteins and spot maps/experimental groups.

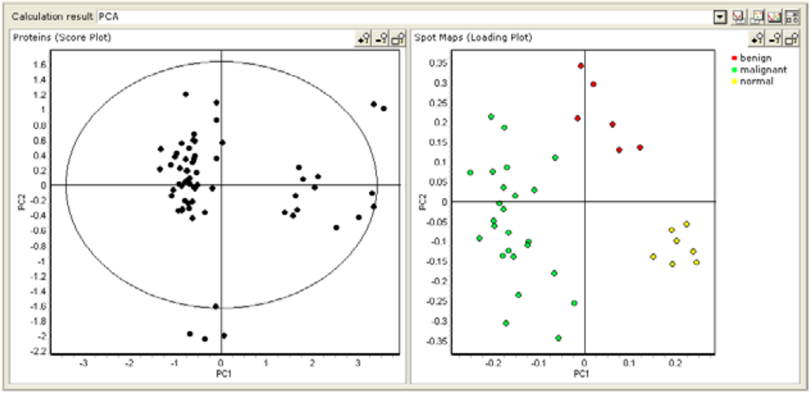
Tip! *For more information on graph settings and zooming, click in the results view (to set the area in focus) and press* ***F1*** *to open the online help for the PCA results view.*

Tip! *To move the plots, right-click on a plot and drag with the mouse.*

8.3.3 Analyze the results of the proteins versus spot maps calculation



The result of the ***proteins versus spot maps*** calculation is presented in the PCA results view. Usually, this analysis has been performed first to get an initial overview of the protein data. The score plot shows proteins and the loading plot spot maps.



When analyzing the results of the proteins versus spot maps PCA calculation:

1 **Start by looking at the left plot**.
The score plot shows an overview of the proteins. The ellipse represents a 95% significance level. Proteins outside of the ellipse are outliers and should be checked. Protein outliers can be either very strongly differentially expressed proteins or mismatched spots.

Check outliers as follows:

a. Select the protein to check by clicking on the spot in the score plot and on a spot map in the loading plot.

b. Select **Tools:Open Source** in the menu bar.

c. The BVA workspace containing the protein will open (with the chosen protein selected). It is possible to check if the protein is mismatched and, if necessary, re-match the protein.

Note: *If changing the BVA workspace, the EDA workspace is not automatically updated but must be re-created.*

Note: *If the protein outlier is mismatched, it is also possible to exclude it from the set in EDA (instead of re-matching it) to avoid re-creating the EDA workspace. See section 12.1.1, Create a set by selecting data for more information.*

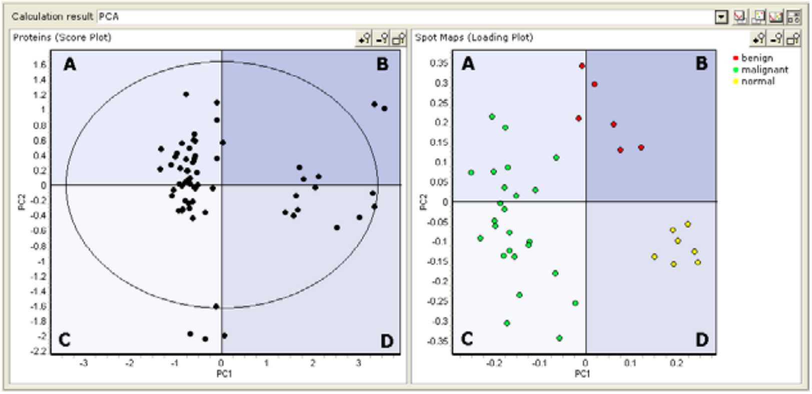
2 **Compare the two plots to see the relationship between the grouping of proteins and spot maps**
It is possible to perform a **rough** comparison of the relationship between proteins and spot maps and to **estimate** which proteins are up- or down-regulated in the different spot maps. Proteins and spot maps located in corresponding quadrants have a connection.

A few examples:

- Proteins in quadrant B and D are probably up-regulated on the spot maps in quadrant B and D and down-regulated on the spot maps in quadrant A and C.

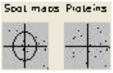
  In the opposite way, proteins in quadrant A and C are probably up-regulated on spot maps in quadrant A and C and down-regulated in spot maps in quadrant B and D.

- Proteins in quadrant B are probably more up-regulated on spot maps in quadrant B than on spot maps in quadrant D (and vice versa).

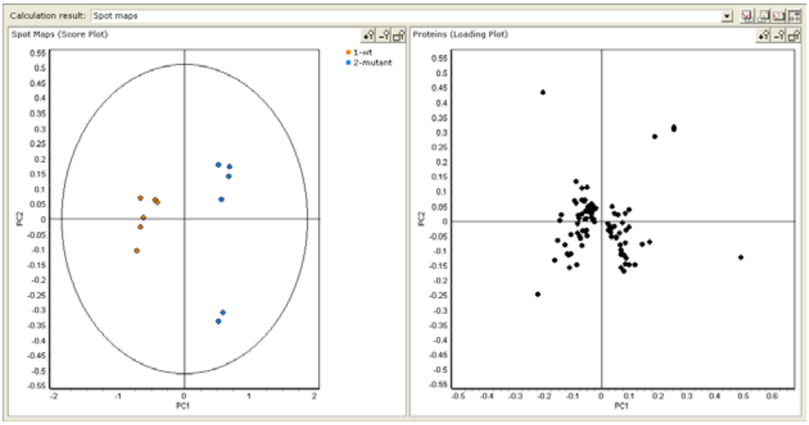


3 For more information and examples on how to interpret the PCA results, see Appendix D.

8.3.4 Analyze the results of the spot maps versus proteins calculation



The result of the ***spot map versus proteins*** calculation is presented in the PCA results view. Usually, this analysis has been performed at an early stage to get an initial overview of the spot map data. The score plot shows spot maps and the loading plot proteins.


Calculation result: Spot maps

When analyzing the results of the spot maps versus proteins PCA calculation look at the left plot:

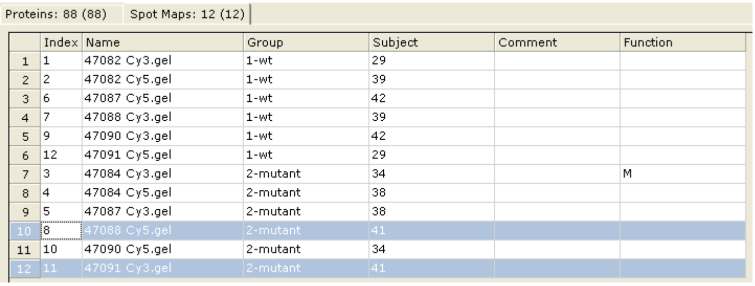
1. The score plot (left plot) shows an overview of the spot maps. The ellipse represents a 95% significance level.

2. The colors for experimental groups are displayed in the plot and a color legend with group names is displayed in the top-right corner.

   Note: *The colors for experimental groups are set in the* ***Setup*** *window of EDA. See section 5.3.6, Edit experimental groups for information on how to edit the color for a group.*

3. Spot maps belonging to the same group should be grouped together. If they are not, this indicates that something is wrong with the spot map, e.g. it contains mis-matched proteins, or if using biological replicates, possibly one individual responds differently to a treatment than the rest of the individuals (if the spot map deviating from the rest belongs to a treated group).

   In the example above, two spot maps in the 2-mutant group deviate from the rest of the spot maps in the group. In this case the two spot maps belong to the same biological replicate, indicating that this biological replicate differs from the other two (can be viewed by selecting the spot maps and checking the subject for the spot maps in the Spot map table).



Proteins: 88 (88) Spot Maps: 12 (12)

	Index	Name	Group	Subject	Comment	Function
1	1	47082 Cy3.gel	1-wt	29		
2	2	47082 Cy5.gel	1-wt	39		
3	6	47087 Cy5.gel	1-wt	42		
4	7	47088 Cy3.gel	1-wt	39		
5	9	47090 Cy3.gel	1-wt	42		
6	12	47091 Cy5.gel	1-wt	29		
7	3	47084 Cy3.gel	2-mutant	34		M
8	4	47084 Cy5.gel	2-mutant	38		
9	5	47087 Cy3.gel	2-mutant	38		
10	8	47088 Cy5.gel	2-mutant	41		
11	10	47090 Cy5.gel	2-mutant	34		
12	11	47091 Cy3.gel	2-mutant	41		

8.3.5 Analyze the results of the proteins versus experimental groups calculation

The results are analyzed in the same way as in the ***protein versus spot maps*** calculation. The difference from the ***protein versus spot maps*** calculation is that the loading plot shows experimental groups instead of spot maps, where each protein's expression in an experimental group is calculated as the mean of that protein's expression on all spot maps in the experimental group.

In the case of many experimental groups and many spot maps in each experimental group, it can be easier to see the relation between proteins and experimental groups in this analysis.

Note: *Before performing this analysis, it should be checked that no spot map outliers exist by performing the **spot maps versus protein** calculation.*

8.3.6 Analyze the results of the experimental groups versus proteins calculation

The results are analyzed in the same way as in the ***spot maps versus proteins*** calculation. The difference from the ***spot maps versus proteins*** calculation is that the score plot shows experimental groups instead of spot maps, where each protein's expression in an experimental group is calculated as the mean of that protein's expression on all spot maps in the experimental group.

In the case of many experimental groups and many spot maps in each experimental group, it can be easier to see the grouping of experimental groups.

Note: *Before performing this analysis, it should be checked that no spot map outliers exist by performing the **spot maps versus protein** calculation.*

9 Calculation and Results - Pattern Analysis

9.1 Introduction

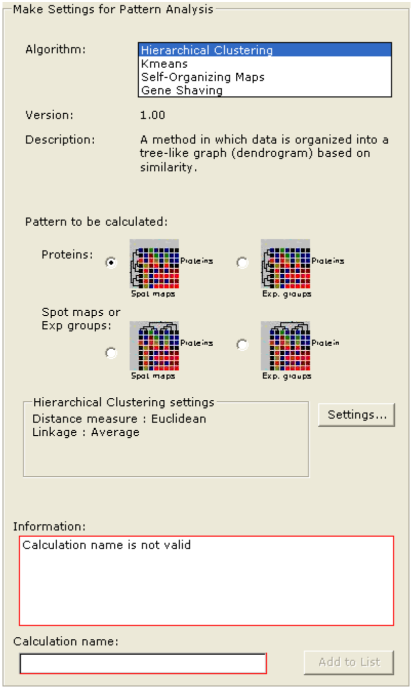
This chapter gives an overview of how to:

- Select settings for the different analyses in the **Make settings for Pattern Analysis** area of the **Calculations** window
- Analyze the results for the pattern analyses in the **Results** window

9.2 Overview

One way to visualize and organize data is to try to group similar data into groups. The Pattern Analysis or Unsupervised Clustering in EDA consists of algorithms that can help to find the subsets of the data (clusters) that show similar expression patterns.

The settings for Pattern Analysis include selecting what types of pattern analysis to perform, what type of pattern to calculate and, if required, changing the settings for the selected pattern analysis algorithm.



Make Settings for Pattern Analysis

Algorithm: Hierarchical Clustering
Kmeans
Self-Organizing Maps
Gene Shaving

Version: 1.00

Description: A method in which data is organized into a tree-like graph (dendrogram) based on similarity.

Pattern to be calculated:

Proteins: ● ○

Spot maps or Exp groups: ○ ○

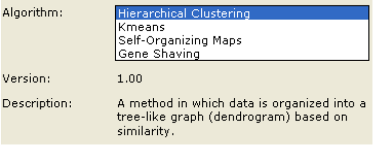
Hierarchical Clustering settings
Distance measure : Euclidean
Linkage : Average

Information:

Calculation name is not valid

Calculation name:

Four different types of pattern algorithms are available: **Hierarchical clustering**, **K-means**, **Self-Organizing Maps (SOM)** and **Gene Shaving**.If you want to add your own algorithms, contact GE Healthcare.



Algorithm: Hierarchical Clustering
Kmeans
Self-Organizing Maps
Gene Shaving

Version: 1.00

Description: A method in which data is organized into a tree-like graph (dendrogram) based on similarity.

Table 9-1 lists examples of when the different analyses may be selected. See Appendix E, Statistics and algorithms - Pattern Analysis for more information on the different algorithms.

Biological query	**Analysis to select**	**See:**
Without having any known parameters, find proteins that co-vary.	Hierarchical clustering	Sections 9.3 and 9.4.
Find proteins that vary in similar ways and place them into a defined number of clusters.	K-means clustering (partitioning clustering)	Sections 9.5 and 9.6.
Find proteins that vary in similar ways and place them into a defined number of clusters but keep the topology of the data, i.e. clusters that show similar profiles are shown next to each other.	Self Organizing Maps (SOM, (partitioning clustering))	Sections 9.5 and 9.6.
Find the most homogenous proteins and exclude small clusters to view only the major clusters.	Gene Shaving (partitioning clustering)	Sections 9.5 and 9.6.

Table 9-1. Pattern analysis methods and recommendations.

Data can be grouped in two dimensions. If performing pattern analysis on proteins, proteins with similar expression patterns will be placed in the same group. It is then also possible, for example, to perform pattern analysis on spot maps. Spot maps where the overall protein expression is similar will be placed in the same group (for example replicate spot maps). In hierarchical clustering, the two-dimensional grouping can be viewed in a dendrogram.

9.3 Make settings for hierarchical clustering

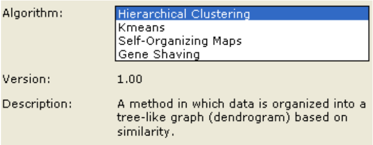
Hierarchical clustering is a method that combines or splits the data pairwise and thereby generates a treelike structure called a dendrogram. The analysis gives an overview of the data by re-arranging the data set into a new, better ordered data set.

It is recommended to start the pattern analysis by performing a two-dimensional hierarchical clustering of proteins and spot maps.

Usually, the Hierarchical clustering analysis should be performed on a smaller set than the base set, containing proteins extracted from the differential expression analysis.

To select settings:

1 In the **Algorithm** area, select the pattern analysis **Hierarchical clustering**.

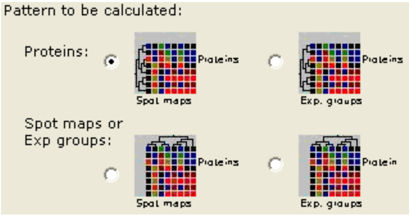


Algorithm: Hierarchical Clustering

Version: 1.00

Description: A method in which data is organized into a tree-like graph (dendrogram) based on similarity.

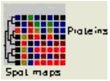
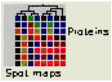
2 In the **Pattern to be calculated** area, select what type of pattern to calculate. See Table 9-2 for information about the patterns.



Pattern to be calculated:

Proteins: ● Spot maps/Proteins

Tip! *It is recommended to add the two corresponding calculations in the Proteins and Spot maps and exp groups areas (**Proteins - Spot maps** and **Spot maps - Proteins** or **Proteins - Experimental groups** and **Experimental groups - Proteins)** to the calculation list to obtain a two-dimensional clustering of the data. For each pattern, a separate calculation must be added to the calculation list*.

Pattern to be calculated	Description
Proteins - Spot maps* 	Select this pattern to cluster the proteins in the data set based on the protein expression from the spot maps. Proteins with similar expression profiles (i.e. similar expression over the spot maps) will be clustered together.
Spot maps - Proteins* 	Select this pattern to cluster the spot maps in the data set. Spot maps with similar overall protein expression (e.g. replica spot maps, spot maps in the same experimental group) will be clustered together.
Proteins - Exp. groups** Proteins Exp. groups	Select this pattern to cluster the proteins in the data set based on the protein expression in the experimental groups. Proteins with similar expression profiles (i.e. similar expression over the expression groups) will be clustered together.
Exp. groups - Proteins** Protein Exp. groups	Select this pattern to cluster the experimental groups in the data set. Experimental groups with similar overall protein expression will be clustered together.

* It is recommended to start with a two-dimensional clustering of proteins and spot maps.

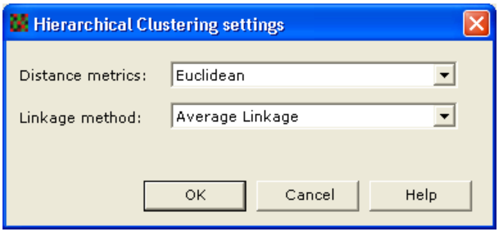
** The protein expression for an experimental group is calculated as the mean of the protein's expression on the spot maps in the experimental group. Therefore, it is a good idea to check that no spot map outliers exist in the different experimental groups by performing PCA on spot maps or by calculating the Hierarchical clustering pattern for Spot maps - Proteins before calculating this pattern.

Table 9-2. Hierarchical clustering patterns.

3 The **Hierarchical Clustering Settings** area shows the default settings for the hierarchical clustering algorithm. These settings can normally be used.

If you want to change the settings, click the **Settings** button. The **Hierarchical Clustering Settings** dialog opens. Change the settings as required and click **OK**.

See Appendix E, Statistics and algorithms - Pattern Analysis for information about the settings for hierarchical clustering.



Hierarchical Clustering settings

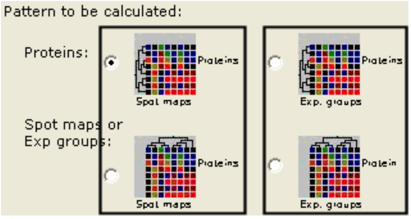
Distance metrics: Euclidean

Linkage method: Average Linkage

4 Use the default name for the calculation in the **Calculation name** field and click **Add to List**.

The calculation is added to the calculation list.

5 Repeat step 2-4 to add the corresponding calculation so that a two-dimensional clustering is obtained (recommended).



Pattern to be calculated:

Proteins: ● ○

Spot maps or Exp groups: ○ ○

Note: *Only one calculation for protein clustering and one calculation for spot maps/experimental groups clustering per set can be added. If adding more, a dialog will appear, asking if you want to overwrite the corresponding previous analysis.*

6 Click **Calculate** to perform the calculation

or

Add other types of calculations (PCA, Pattern Analysis and Discriminant Analysis) to the **Calculation** list (see Chapter 6 for information about the workflow).

7 When a calculation has finished, this will be indicated by a status icon in front of the calculation. The following status icons may appear in front of the calculations:

Icon	Description
🟢	The calculation is in progress.
✓	The calculation has successfully finished.
✗	The calculation has been cancelled.
🔴	The calculation has failed.

The status of the calculations will also be displayed in the **Calculation status** field.

For information on how to analyze the results, see section 9.4.

9.4 Analyze the results of the hierarchical clustering

9.4.1 Overview of the results

The results of the hierarchical clustering are displayed in the form of one or two dendrograms (depending on the calculations performed) together with the heat map (see section 3.3 for detailed information about the heat map).

The dendrogram orders the data so that similar data is displayed next to each other. It is possible to see which proteins and/or spot maps/experimental groups have been grouped together at each step of the algorithm. Proteins with similar expression profiles are grouped together and spot maps/experimental groups with similar overall protein expression (e.g. replica spot maps) are grouped together.

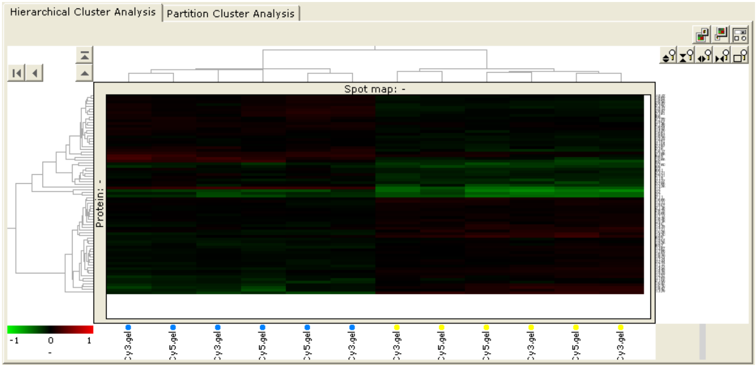


Fig 9-1. Results of a two-dimensional clustering of proteins and spot maps.

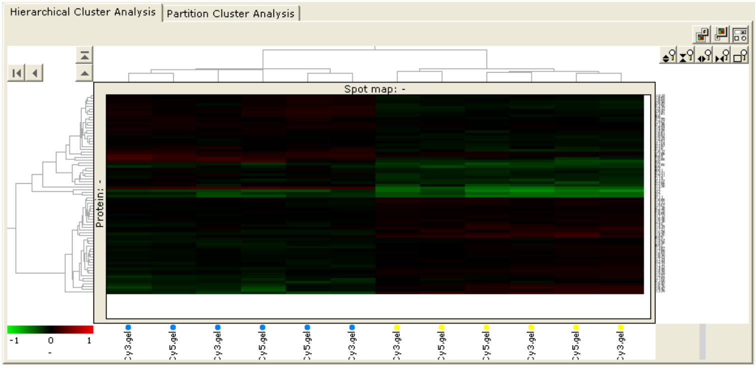
When analyzing the hierarchical clustering, the same principles apply independent of which calculations have been performed. Either *proteins and/or spot maps* are clustered or *proteins and/or experimental groups* are clustered.

9.4.2 Analyze clustering of proteins and spot maps

1 Select **Results** in the workflow area and then Pattern Analysis in the Results bar to display the results of the hierarchical clustering.

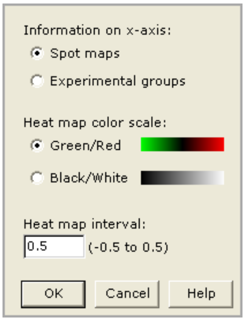
2 The results are displayed by default (otherwise select the **Hierarchical Cluster Analysis** tab).

Depending on the calculations performed, dendrograms for clustering of *proteins and/or spot maps* or *proteins and/or experimental groups* are displayed.



3 The default value for the heat map interval is set to 1. If this setting gives weak signals in the heat map, change the heat map interval as follows:

a. Click the **Settings** icon to display the **Heat map settings** pop-up dialog.


Information on x-axis:

● Spot maps

○ Experimental groups

Heat map color scale:

● Green/Red

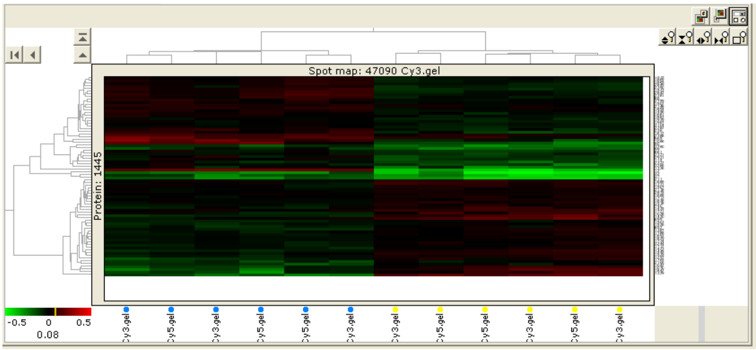
○ Black/White

Heat map interval:

0.5 (-0.5 to 0.5)

OK Cancel Help

b. Change the heat map interval to, for example, 0.5 and click **OK**. The heat map is updated.



4 Analyze the results as follows (in this case a two-dimensional clustering of

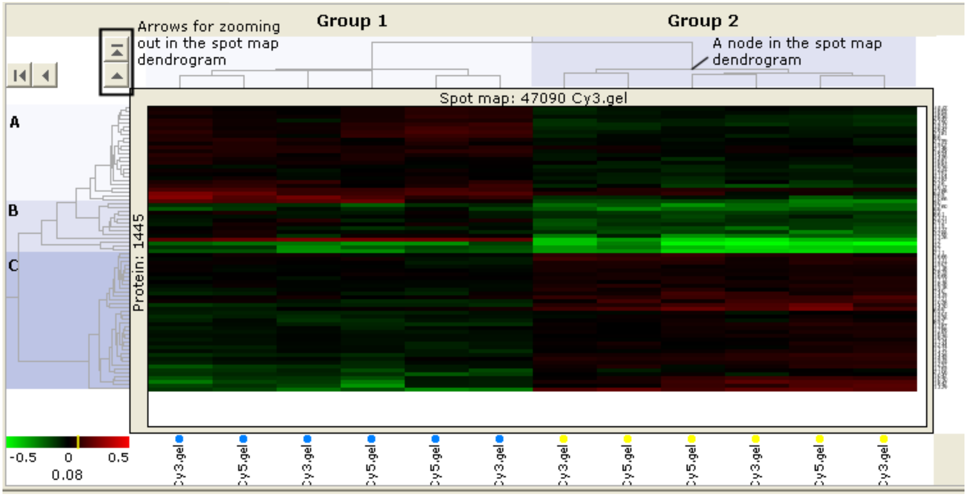
proteins and spot maps):

- *View which spot maps have been clustered together*
  Analyze clustering of spot maps to check that replicate spot maps are grouped together. If they are not, this indicates that something is wrong with the spot map For example, it may contain mis-matched proteins, or if biological replicates were used, one individual may respond differently to a treatment than the rest of the individuals (if the spot map deviating from the rest belongs to a treated group).

  In the example below, the spot maps have been divided into two clusters. This can be determined by looking at the dendrogram for the spot maps and the color coding of the spot maps for the two groups (blue and yellow)

  To zoom in on the spot map dendrogram, double-click a node in the dendrogram to display only the spot maps clustered by the node in the heat map and in the spot map table. To zoom out, use the arrows at the top left corner. For more information about the Results view, click the heat map and press **F1** to open the online help for the hierarchical clustering results view

  For information on zooming in the heat map, see section 3.3.2, Zooming within the heat map.



- *A rough estimation of the number of protein groups with the same expression patterns can be made*
  It is possible to see the main groups of proteins in the protein dendrogram.

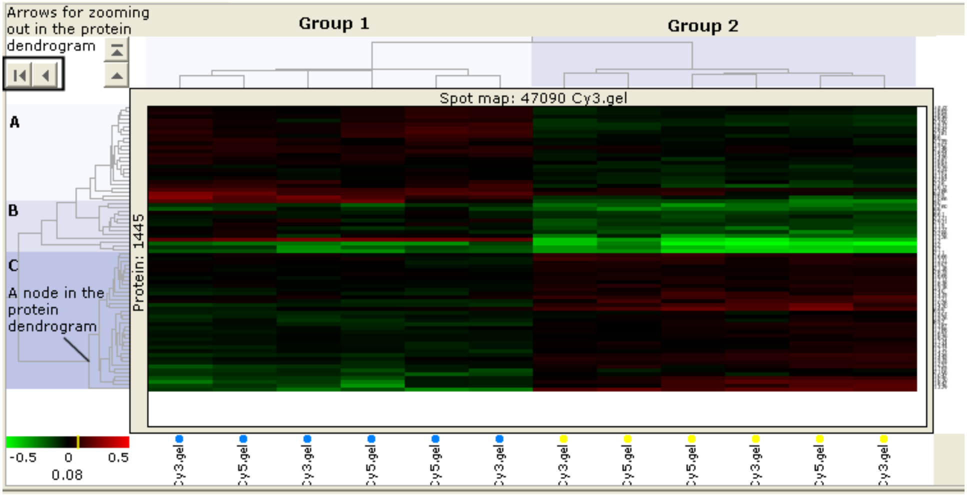
  In the example below, approximately three main groups of proteins can be seen: A, B and C, although some proteins deviate from their group.

  In group A, most proteins are up-regulated in Group 1 and down-regulated in Group 2.

  In group B, most proteins are down-regulated in both groups but are more strongly down-regulated in Group 2.

  In group C, most proteins in Group 1 are down-regulated and up-regulated in Group 2.

Note: *To obtain a more detailed grouping of proteins with the same expression profiles, perform partitioning clustering of proteins, for example K-means clustering.*



9.4.3 Analyze clustering of proteins and experimental groups

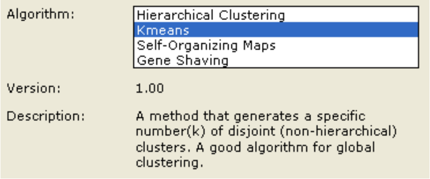
The results are analyzed in the same way as in the ***proteins versus spot maps*** calculation. The difference from the ***proteins versus spot maps*** calculation is that the heat map shows experimental groups instead of spot maps, where each protein's expression in an experimental group is calculated as the mean of that protein's expression on all spot maps in the experimental group.

Therefore, grouping of spot maps cannot be viewed, only the experimental groups that have similar overall expression. Grouping of proteins are viewed as in the ***proteins versus spot maps*** calculation.

Note: *Analyze clustering of experimental groups only if you already know from previous analyses that the replicate spot maps are similar, otherwise important information may be lost.*

9.5 Make settings for partitioning clustering

1 In the **Algorithm** area, select the pattern analysis to calculate (**K-means**, **Self-Organizing Map** or **Gene Shaving**). See Table 9-1 for information on pattern analyses.



Algorithm: Hierarchical Clustering
Kmeans
Self-Organizing Maps
Gene Shaving

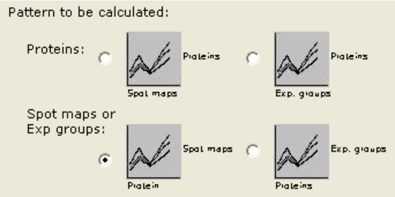
Version: 1.00

Description: A method that generates a specific number(k) of disjoint (non-hierarchical) clusters. A good algorithm for global clustering.

2 In the **Pattern to be calculated** area, select what type of pattern to calculate.

Tip! *Usually, one is interested in clustering the proteins to determine the number of protein clusters and the cluster expression profiles.*

See Table 9-3 for information about the patterns.



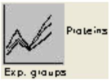
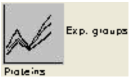
Pattern to be calculated:

Proteins: ○ ○

Spot maps or Exp groups: ● ○

For each pattern to be calculated, a separate calculation must be added to the calculation list (e.g. set up one calculation that calculates the pattern for proteins and set up another calculation that calculates the pattern for spot maps).

Note: *As many calculations as required can be added. Enter appropriate names for the calculations*

Pattern to be calculated	Description
Proteins - Spot maps 	Select this pattern to cluster the proteins in the data set based on the protein expression from the spot maps. Proteins with similar expression profiles (i.e. similar expression over the spot maps) will be clustered together. In the results view, each cluster will be displayed in a separate graph.
Spot maps - Proteins 	Select this pattern to cluster the spot maps in the data set. Spot maps with similar overall protein expression (e.g. replica spot maps, spot maps in the same experimental group) will be clustered together. In the results view, each cluster will be displayed in a separate graph.
Proteins - Exp. groups* 	Select this pattern to cluster the proteins in the data set based on the protein expression in the experimental groups. Proteins with similar expression profiles (i.e. similar expression over the expression groups) will be clustered together. In the results view, each cluster will be displayed in a separate graph.
Exp. groups - Proteins* 	Select this pattern to cluster the experimental groups in the data set. Experimental groups with similar overall protein expression will be clustered together. In the results view, each cluster will be displayed in a separate graph.

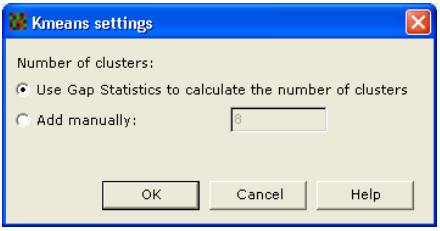
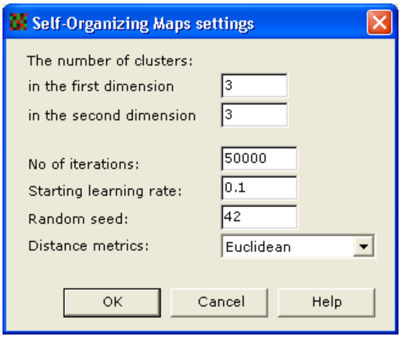
* The protein expression for an experimental group is calculated as the mean of the protein's expression on the spot maps in the experimental group. Therefore, it is a good idea to check that no spot map outliers exist in the different experimental groups by performing PCA on spot maps or by calculating the Hierarchical clustering pattern for Spot maps - Proteins before calculating this pattern.

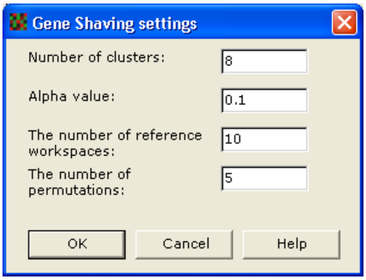
Table 9-3. Partitioning clustering patterns.

3 The **Settings** area shows the default settings for the currently selected algorithm. These settings can normally be used.

If you want to change the settings, click the **Settings** button. The **Settings** dialog for the currently selected analysis opens. Change the settings as required and click **OK**.

See Appendix E, Statistics and algorithms - Pattern Analysis for information about the settings for the different analyses. The table below lists the different **Settings** dialogs showing the default settings. Tips on some of the settings are also displayed.

Analysis	Default settings
K-means clustering	***Tip!*** *If the number of clusters to be achieved after grouping is known, select the **Add manually** radio button and enter the number in the field to the right.* ***Tip!*** *If a hierarchical clustering has already been performed, it is possible to estimate the number of clusters in the data by viewing the dendrogram.*  Kmeans settings Number of clusters: ● Use Gap Statistics to calculate the number of clusters ○ Add manually: 8
SOM	***Tip!*** *The number of clusters is determined by the product of X and Y. However, if the actual number of clusters is less, some of the clusters will contain zero proteins.* ***Tip!*** *If a hierarchical clustering has already been performed, it is possible to estimate the number of clusters in the data by viewing the dendrogram.*  Self-Organizing Maps settings The number of clusters: in the first dimension: 3 in the second dimension: 3 No of iterations: 50000 Starting learning rate: 0.1 Random seed: 42 Distance metrics: Euclidean

Analysis	Default settings
Gene Shaving	***Tip!*** *If the number of clusters to be achieved after grouping is known, select the* ***Add manually*** *radio button and enter the number in the field to the right.* ***Tip!*** *If a hierarchical clustering has already been performed, it is possible to estimate the number of clusters in the data by viewing the dendrogram.*  Gene Shaving settings Number of clusters: 8 Alpha value: 0.1 The number of reference workspaces: 10 The number of permutations: 5

4 Enter a name for the calculation in the **Calculation name** field and click **Add to List**.

The calculation is added to the calculation list.

5 Click **Calculate** to perform the calculation

or

Add more pattern analysis calculations or other types of calculations to the **Calculation** list (see Chapter 6 for information about the workflow).

Note: *As many partitioning clustering calculations as required can be added to the calculation list.*

6 When a calculation has finished, this will be indicated by a status icon in front of the calculation. The following status icons may appear in front of the calculations:

Icon	Description
🟢	The calculation is in progress.
✓	The calculation has successfully finished.
✗	The calculation has been cancelled.
🔴	The calculation has failed.

7 The status of the calculations will also be displayed in the **Calculation status** field.

For information on how to analyze the results, see section 9.6.

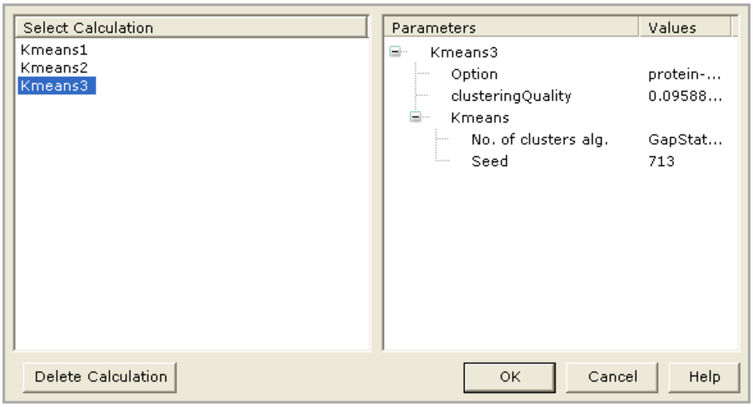
9.6 Analyze partitioning clustering

The results of the possible partitioning clustering methods are displayed in the Partition Cluster Analysis tab in the **Results** window for pattern analysis.

9.6.1 Select the calculation from which to view the results

1 Select the **Results** step in the workflow area, Pattern Analysis in the Results bar and then the **Partition Cluster Analysis** tab. The results for the partition clustering calculation in the **Calculation result** field are displayed in the results view.

2 If you want to view the results for another partition clustering calculation (if several were performed), click the **Calculation result** arrow button to display the **Select calculation** pop-up dialog.



Select Calculation
Kmeans1
Kmeans2
Kmeans3

Parameters	Values
Kmeans3	
Option	protein-...
clusteringQuality	0.09588...
Kmeans	
No. of clusters alg.	GapStat...
Seed	713

3 Select the calculation from which to display the results in the **Select calculation** column. The values for the parameters in the calculation are shown in the area to the right (**Parameters** and **Values** columns).

4 Click **OK** to display the results for the selected calculation in the results view.

9.6.2 Analyze the results for partitioning clustering of proteins

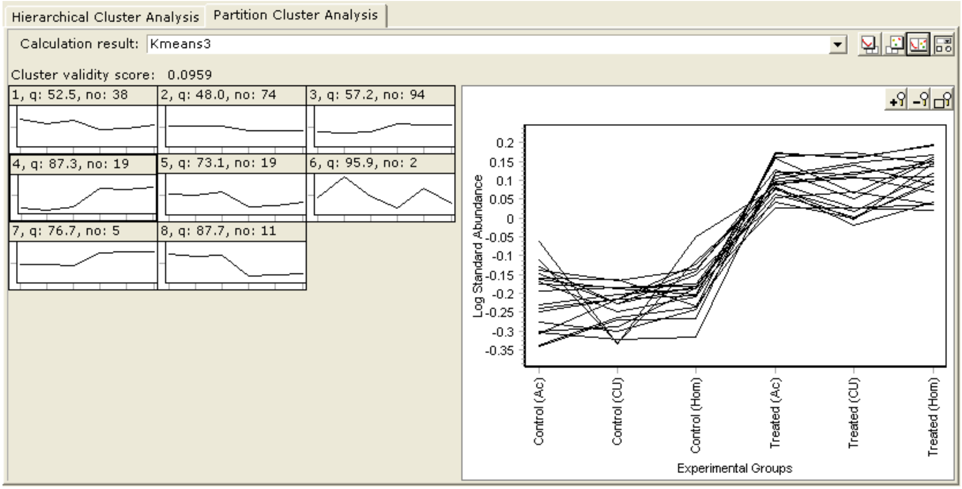
The results of the K-means, SOM and Gene Shaving clustering calculations are analyzed in the same way. See Table 9-1 for information on the differences between the calculations and how data is clustered.

The analysis provides information on the number of protein clusters in the data and the expression profiles in the clusters.

To analyze the results:

1 In the **Partition Cluster Analysis** tab, select the calculation for which results to view in the **Calculation** result field.

2 The results are displayed in the Results view.



Calculation result: Kmeans3

Cluster validity score: 0.0959

3 The left view shows the clusters calculated by the algorithm. Each cluster contains proteins with the same expression profile. Two quality parameters are displayed:

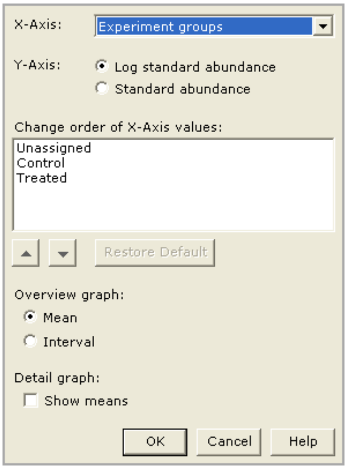
- The **Cluster validity score** measures the quality of the clustering. This score can be used to compare the quality of the different clusterings performed. The higher the cluster validity score, the better the clustering.
- For each cluster, a quality measure (**q**) and the number of proteins in the cluster are displayed.
  The q-value is a number between 1 and 100 and measures the homogenity of a cluster. If the expression pattern for the proteins in a cluster is identical, the value will be 100. It is not possible to compare the q-values for different clustering analyses.

4 The right view shows the cluster selected in the left view in a detailed graph.

Click on the different clusters in the left view to display the cluster in the right view.

5 If you want to change settings for the graphs in the left and right views, click the **Settings** icon in the right view.

The **Partition clustering graph pop-up** dialog opens.



X-Axis: Experiment groups

Y-Axis: ● Log standard abundance ○ Standard abundance

Change order of X-Axis values:

Unassigned
Control
Treated

Overview graph:
● Mean
○ Interval

Detail graph:
□ Show means

6 Click the **Help** button to open the online help for this dialog and obtain detailed information on the different settings.

7 Edit the settings as appropriate and click **OK**.

9.6.3 Analyze the results for partitioning clustering of spot maps/ experimental groups

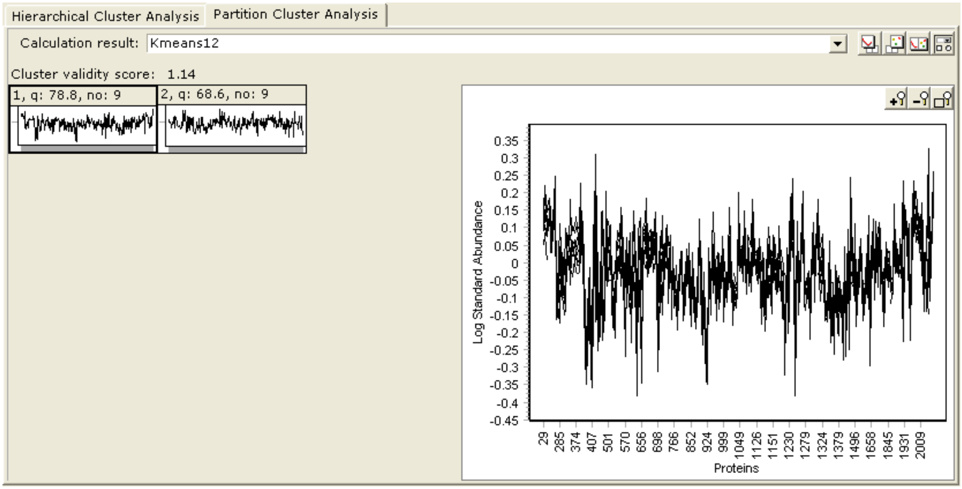
The results of the K-means, SOM and Gene Shaving clustering calculations are analyzed in the same way. See Table 9-1 for information on the differences between the calculations and how data is clustered.

In the analysis, it is possible to see which spot maps/experimental groups were clustered together. In the case of spot maps, the replica spot maps should be grouped together.

To analyze the results:

1 On the **Partition Cluster Analysis** tab, select the calculation from which to view the results in the **Calculation** result field.

2 The results are displayed in the Results view (in this case clustering of spot maps).



Calculation result: Kmeans12

Cluster validity score: 1.14

1, q: 78.8, no: 9

2, q: 68.6, no: 9

3 The left view shows the clusters calculated by the algorithm.

Each cluster contains spot maps (or experimental groups) with the same overall protein expression profile. Typically, replica spot maps should be clustered together. Two quality parameters are displayed:

- The **Cluster validity score** measures the quality of the clustering. This score can be used to compare the quality of the different clusterings performed. The higher the cluster validity score, the better the clustering.
- For each cluster, a quality measure (**q**) and the number of spot maps (or experimental groups) in the cluster are displayed. The q-value (0-100)

measures the homogenity of a cluster where 100 means that the spot maps have identical overall protein expression profiles.

4 The right view shows the cluster selected in the left view in a detailed graph. Click on the different clusters in the left view to display the cluster in the right view.

5 In the example on the previous page, two defined clusters with spot maps were found, corresponding to the experimental setup with two experimental groups.

10 Calculation and Results - Discriminant Analysis

10.1 Introduction

This chapter gives an overview of how to:

- Select settings for the different analyses in the **Select settings** area (middle panel) of the **Calculations** window
- Analyze the results for the different types of analyses in the **Results** window

An overview of the chapter content is outlined in the table below:

Information on:	**See:**
Settings and analysis - Overview	Section 10.2
Workflow	Section 10.3
Select settings for the Marker Selection calculation	Section 10.4
Analyze the results of the Marker Selection calculation	Section 10.5
Select settings for the Classifier Creation calculation	Section 10.6
Analyze the results of the Classifier Creation calculation	Section 10.7
Select settings for the Classification calculation	Section 10.8
Analyze the results of the Classification calculation	Section 10.9

10.2 Settings and analysis - Overview

The discriminant analysis calculation consists of three parts: marker selection, classifier creation and classification.

The ***Marker Selection*** analysis can be used to find a set of proteins that can be used to discriminate between experimental groups, e.g. benign tumors and malignant tumors.

If such a set is found, it is possible to create a classifier specialized for discriminating between e.g. the benign tumors and malignant tumors experimental groups (***Classifier Creation***).

Once a classifier has been created, it can be used to classify (***Classification***) a new data set of spot maps to the correct experimental groups.

Note: *To be able to find features/biomarkers in a data set, the data must have been classified by another method, for example by clinical diagnosis of the samples in the case of benign and malignant tumors, and the spot maps must have been placed in the correct groups.*

Note: *It is important to have a balanced workspace when performing discriminant analysis in order to obtain good results. This means that the groups used in marker selection and classifier creation should have approximately the same number of spot maps.*

10.3 Workflow

An example of an overall workflow for discriminant analysis is outlined below:

1 **Find markers**
If you need to create a classifier or just want to find features/biomarkers, start with the calculation **Marker Selection**. Select the experimental groups of known class and set up one or more variants of the calculation. Add the calculation(s) to the calculation list and calculate.

Note: *If you already have a classifier and want to classify a new data set, go straight to the Classification calculation (step 5).*

See section 10.4 for more information.

2 **View the results of the Marker Selection calculation**
Select the **Results** step to view the results of the **Marker Selection** calculation and to determine which protein set best discriminates between the selected experimental groups. Then create a new set with these proteins.
See section 10.5 for more information.

3 **Create a classifier**
Go back to the **Calculations** step. Select the **Classifier Creation** calculation, select the created set in step 2, the experimental groups of known classes and enter settings for creating a classifier. One or more calculations with different settings can be created and added to the calculation list to evaluate different classifiers.
See section 10.6 for more information.

4 **View the results of the classifier(s)**
Select the **Results** step to view the results of the **Classifier Creation** calculation(s) and to determine which classifier performs best if several were created.
See section 10.7 for more information.

5 **Classify**
Go back to the **Calculations** step and select the **Classification** calculation. Select a set containing the unknown data which is to be classified and choose a classifier. Add the calculation to the calculation list and calculate.
See section 10.8 for more information.

6 **View the results of the classification**
Select the **Results** step to view the results of the **Classification** calculation. The results of the classification of the spot maps are displayed in the Spot Map table.
See section 10.9 for more information.

10.4 Make settings for the Marker Selection calculation

10.4.1 Overview

The **Marker Selection** method is used to find a set of proteins that can be used to discriminate between different properties of the data, for example benign tumors and malignant tumors.

Note: *It is possible to use all data in a data set and go straight to the Classifier Creation method, but usually a smaller set of proteins can be used to classify the data with the same or even better accuracy than using all of the proteins in the data set. It can also be of importance to find a smaller set of proteins in order to identify biomarkers.*

To find features, the data must already have been classified by a different method, for example by diagnosis. The property discriminating between classes of data can be an experimental group or a condition.

The finding of features is divided into 4 processes:

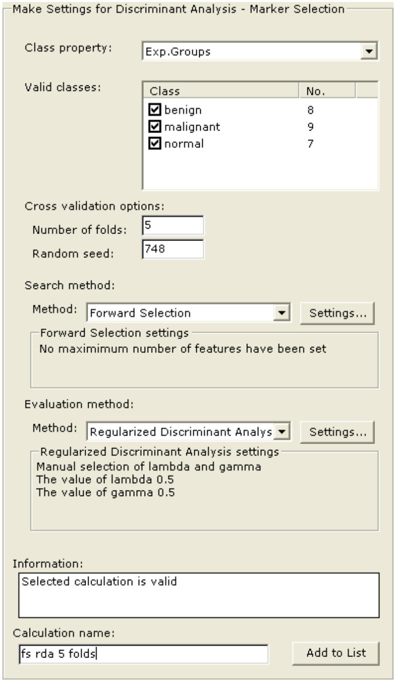
- **Defining the property to use for discrimination**
  Select which property to use for discrimination (experimental group or condition).

- **Setting up cross validation options**
  This is performed to establish how to divide the data (i.e. into how many number of folds) used for searching and evaluation of protein sets. For example, if dividing the data into five folds (parts), fold 1-4 will be used in the search method and fold 5 for evaluation. Then fold 2-5 will be used in the search method and fold 1 for evaluation and so on.

- **Selecting search method**
  The search method searches for proteins that best discriminate between the classes according to the test options and ranks the proteins and protein sets.

- **Selecting evaluation method**
  The evaluation method evaluates the different protein sets found by the search method. It classifies the data using the different protein sets (found by the search method) and compares the different results with the real data (where class is known) according to the cross validation options. In the results step, the results are displayed in a graph with proteins on the x-axis and accuracy of class determination (in %) displayed on the y-axis. The number of created classifiers will be the same as the number of folds entered in the cross validation options.

10.4.2 Make settings



Make Settings for Discriminant Analysis - Marker Selection

Class property: Exp.Groups

Valid classes:

Class	No.
☑ benign	8
☑ malignant	9
☑ normal	7

Cross validation options:

Number of folds: 5

Random seed: 748

Search method:

Method: Forward Selection

Forward Selection settings

No maximimum number of features have been set

Evaluation method:

Method: Regularized Discriminant Analys

Regularized Discriminant Analysis settings

Manual selection of lambda and gamma

The value of lambda 0.5

The value of gamma 0.5

Information:

Selected calculation is valid

Calculation name: fs rda 5 folds

1 Select a property, i.e. experimental group or condition, that defines the different classes of data from the **Class property** drop-down list.

All classes for the property in the selected set, i.e. all experimental group names or all condition names, are shown in the **Valid classes** field.

2 Select the experimental groups/ conditions with known class by checking the appropriate boxes in the **Valid classes** field.

3 In the **Cross validation option** area:

a. Select the number of folds to use in the marker selection. The default value of 5 can be used in most cases. This means that the data will be divided into five parts. However, if any of the valid classes contains less than 5 spot maps, the number of folds should be decreased, so that the number of folds is <= the number of spot maps in the experimental group with the least spot maps. See Appendix F, Statistics and algorithms - Discriminant Analysis for more information on folds.

b. Use the default value in the **Seed** field. If you want to repeat an experiment and obtain exactly the same results, enter the same Seed value as the one used for the repeat experiment (all other parameters in the calculation must also be the same).

4 In the **Search method** area, select the search method (**Forward selection** or **Partial least squares**) to use for the searching and ranking of proteins. The forward selection method is the default method.

Note: *It is possible to create two or more different calculations where different search methods and search method settings are selected to test which one gives the best accuracy.*

If you want to change the settings for the selected method, click the **Settings** button. For more information about the two search methods and settings, see Appendix F.

5 In the **Evaluation method** area, select the evaluator method (**K-Nearest Neighbor** or **Regularized discriminant analysis**) to use for evaluation of the protein set found in step 4.

Note: *It is possible to create two or more different calculations where different evaluation methods and evaluation method settings are selected to test which one gives the best accuracy.*

6 Enter a name for the calculation in the **Calculation name** field and click **Add to List**.

The calculation is added to the calculation list.

Note: *As many calculations as required can be added to the calculation list*.

7 Click **Calculate** to perform the calculation

or

Add more marker selection calculations by repeating step 1-6.

8 When a calculation has finished, this will be indicated by a status icon in front of the calculation. The following status icons may appear in front of the calculations:

Icon	Description
🟢	The calculation is in progress.
✓	The calculation has successfully finished.
✗	The calculation has been cancelled.
🔴	The calculation has failed.

The status of the calculations will also be displayed in the **Calculation status** field.

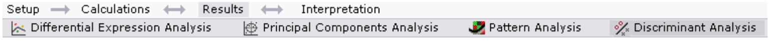
For information on how to analyze the results, see section 10.5.

10.5 Analyze the results of the Marker Selection calculation

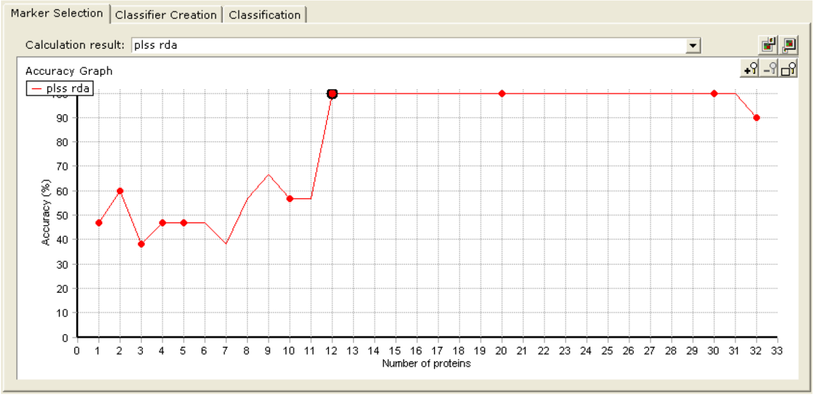
When analyzing the results of the Marker Selection calculation, a set with a minimum of proteins giving the best accuracy when predicting the class should be found. A set with these proteins is then created.

To analyze the results:

1 Select the **Results** step in the workflow area and then **Discriminant Analysis** from the Results bar.



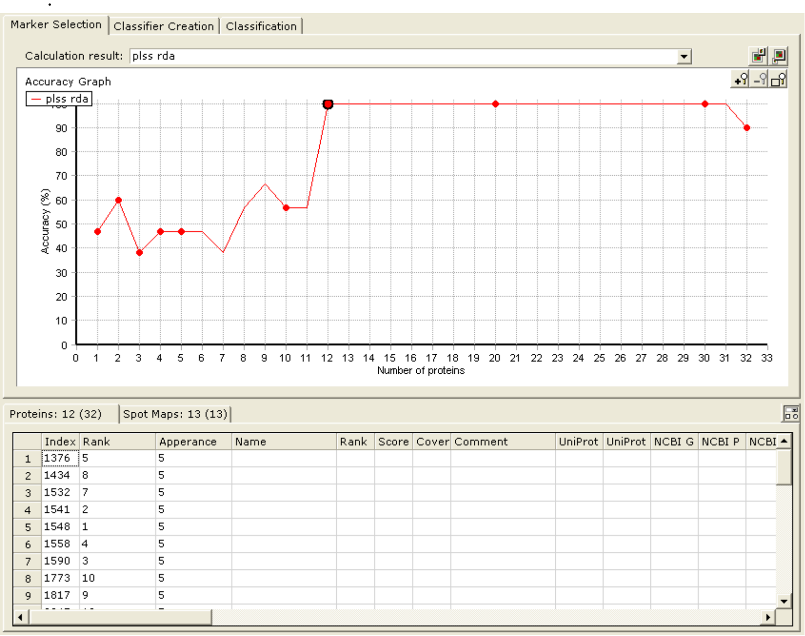
2 The **Results** window is displayed, showing the results of the marker selection calculation in the Results view.



The accuracy graph shows the mean accuracy of class prediction for different numbers of proteins.

Note: *The mean accuracy is calculated using the accuracy from each created classifier. The number of created classifiers is equal to the number of folds used in the calculation.*

3 Select the lowest number of proteins that give the highest accuracy score (preferably 100%) for discriminating between the groups by clicking on this in the accuracy graph. The proteins are shown in the Protein Table.



Calculation result: plss rda

Proteins: 12 (32) | Spot Maps: 13 (13)

	Index	Rank	Apperance	Name	Rank	Score	Cover	Comment	UniProt	UniProt	NCBI G	NCBI P	NCBI
1	1376	5	5										
2	1434	8	5										
3	1532	7	5										
4	1541	2	5										
5	1548	1	5										
6	1558	4	5										
7	1590	3	5										
8	1773	10	5										
9	1817	9	5										

Two parameters indicating the quality of the result are displayed in the Protein Table:

- **Appearance**
  For each protein, the number of classifiers that have selected this protein is listed in the **Appearance** column. If 5 folds were used, 5 in the **Appearance** column means that all classifiers have selected this protein and 1 means that only one classifier has selected the protein. It is primarily this parameter that is used to determine the quality of the results (see step 4).

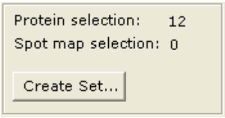
  The number of proteins in the Protein Table is not always identical to the number of proteins that were selected in the Accuracy Graph. The reason for this is that several classifiers are created (the same as the number of folds entered in the Marker Selection calculation) and that the different classifiers do not choose the same proteins as the x best proteins selected in the Accuracy graph.

- **Rank**
  For each protein, a rank value is displayed in the **Rank** column. This value is the mean of the different classifiers' ranking of the protein. Each classifier gives the first protein that is selected by the search method and gives the best result in the evaluation method rank 1, the second protein that is selected receives rank 2 and so on.

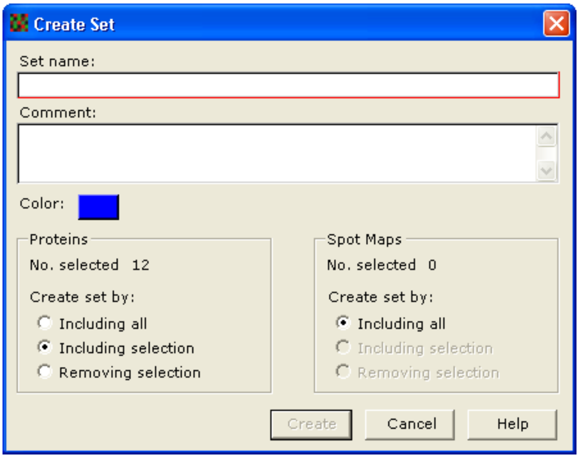
4 If the number of proteins in the Protein Table is about the same as the number of proteins selected in the accuracy graph, i.e. the Appearance value is equal to the number of folds used for most proteins, these proteins are probably good markers and will probably perform well when building a classifier. Continue with step 5.

Note: *If the number of proteins in the Protein Table differs a lot from the number of proteins selected in the accuracy graph, i.e. the Appearance values are low for most proteins (compared to the number of folds), it is recommended to re-run the Marker Selection calculation with another number of folds or other parameters to see if the same proteins appear. Select the proteins that appear in most of the independent marker selection calculations if a result with better appearance values cannot be found.*

5 Select the proteins in the Protein Table and click **Create set...** in the **Set** area. The **Create set dialog** opens.



Protein selection: 12
Spot map selection: 0



Set name:

Comment:

Color: 🟦

Proteins
No. selected 12
Create set by:
○ Including all
● Including selection
○ Removing selection

Spot Maps
No. selected 0
Create set by:
● Including all
○ Including selection
○ Removing selection

6 Enter a name for the set in the **Set name** field.

7 If required, enter a comment on the set in the **Comment** field.

8 The **Including selection** radio button in the **Proteins** area is selected by default. Use this setting.

9 Make sure that the **Including all** radio button in the **Spot Maps** area is selected.

10 Click **Create**. The set is created and will be displayed in the **Select set** field.

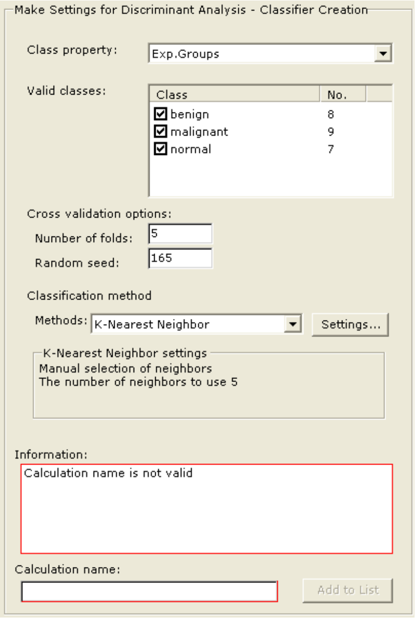
10.6 Make settings for the Classifier Creation calculation

10.6.1 Overview

The **Classifier creation** method is used when the results of the **Marker selection** method have been analyzed and a new set containing the markers that best discriminate between the classes has been created. This set should be used when creating the classifier(s).

10.6.2 Make settings

1 Make sure the set with your markers is selected in the **Select set** field.



Make Settings for Discriminant Analysis - Classifier Creation

Class property: Exp.Groups

Valid classes:

Class	No.
☑ benign	8
☑ malignant	9
☑ normal	7

Cross validation options:

Number of folds: 5

Random seed: 165

Classification method

Methods: K-Nearest Neighbor

K-Nearest Neighbor settings
Manual selection of neighbors
The number of neighbors to use 5

Information:

Calculation name is not valid

Calculation name:

2 Select the property (experimental groups or conditions) that was used in the Marker selection calculation from the **Class property** drop-down list.

All classes for the property in the selected set, i.e. all experimental group names or all condition names, are shown in the **Valid classes** field.

3 Select the experimental groups/ conditions of known class by checking the appropriate boxes in the **Valid classes** field.

4 In the **Cross validation option** area:

a. Enter the number of folds to use in the **Number of folds** field. It is recommended to use the same number as in the Marker selection calculation from which the marker set was created.

b. Use the default value in the **Seed** field.

5 In the classification method area, select the classification method (**K-Nearest Neighbor** or **Regularized discriminant analysis**) to use for classification of the data (it is recommended to use the same method as in the Marker selection calculation from which the marker set was created).

6 Enter a name for the calculation in the **Calculation name** field and click **Add to List**.

The calculation is added to the calculation list.

7 Click **Calculate** to perform the calculation

or

Add more classifier creation calculations by repeating steps 1-6.

Note: *As many calculations as required can be added to the calculation list.*

8 When the calculation(s) are finished, this will be indicated by a status icon in front of the calculation. The status of the calculations will also be displayed in the **Calculation status** field.

For information on how to analyze the results, see section 10.7.

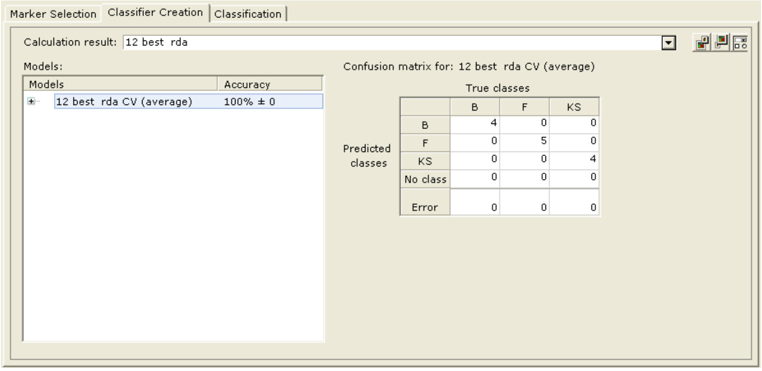
10.7 Analyze the results of the Classifier Creation calculation

Analyze the results of the classifier creation calculation to view how well the created classifier(s) performs.

To analyze the results:

1 Select the **Results** step in the workflow area, select **Discriminant Analysis** in the Results bar and select the **Classifier Creation** tab in the Results view. Then select the appropriate calculation results from the **Calculation result** field.

The results are displayed in the Results view.



Calculation result: 12 best rda

Models:

Models	Accuracy
12 best rda CV (average)	100% ± 0

Confusion matrix for: 12 best rda CV (average)

True classes

		B	F	KS
Predicted classes	B	4	0	0
	F	0	5	0
	KS	0	0	4
	No class	0	0	0
	Error	0	0	0

2 In the **Models** area, the result of the model is presented. It is the **CV (average)** that is the created classifier, but it is possible to see each sub-model that was created at each iteration of the algorithm. The accuracy of class prediction for the created classifier is shown in the **Accuracy** column and the number of wrongly classified spot maps is shown in the **Error** column. The highest accuracy with smallest variation is desirable.

3 In the **Confusion matrix** area to the right a more detailed view of the classification of the spot maps is displayed. Spot maps that were wrongly classified are displayed in red.

Tip! *To get an indication of why a spot map is wrongly classified (for example, if it lies on the border between two groups), select the spot map in the* ***Confusion matrix*** *area and click on PCA in the Results bar. The PCA results for the calculation on the set (if performed) are displayed in the Results view with the wrongly classified spot map selected.*

4 If several classifiers were created, repeat the analysis for each classifier by selecting the result in the **Calculation result** field and performing steps 2-3.

Note which classifier gave the best result and use this classifier in the **Classification** calculation.

10.8 Make settings for the Classification calculation

10.8.1 Overview

Once a classifier has been created, it is possible to analyze a data set with spot maps with unknown class. The available classifiers are displayed in the **Classifiers** list in the settings for the **Classify** method.

Note: *Select a set containing the unknown data (i.e. spot maps with unknown class) when performing this calculation.*

10.8.2 Make settings

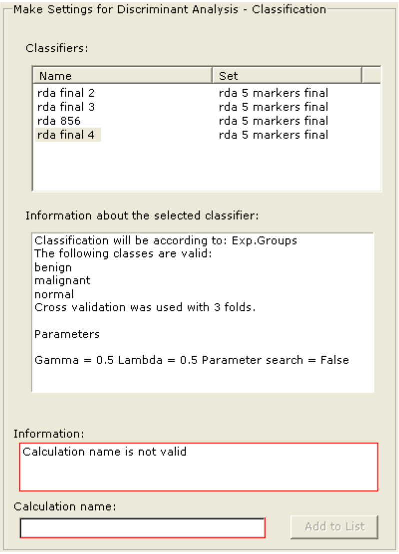
1 Make sure that the set containing the unknown data is selected in the **Select set** field in the **Calculations** area.

2 In the **Classifiers** area select the classifier to use for classification. Information about the selected classifier is shown in the **Information about the selected classifier/model** field.

The classifier to select should be the classifier that gave the best result in the classifier creation calculation.

3 Enter a name for the calculation in the **Calculation name** field and click **Add to List**.

The calculation is added to the calculation list.



Make Settings for Discriminant Analysis - Classification

Classifiers:

Name	Set
rda final 2	rda 5 markers final
rda final 3	rda 5 markers final
rda 856	rda 5 markers final
rda final 4	rda 5 markers final

Information about the selected classifier:

Classification will be according to: Exp.Groups
The following classes are valid:
benign
malignant
normal
Cross validation was used with 3 folds.

Parameters

Gamma = 0.5 Lambda = 0.5 Parameter search = False

Information:

Calculation name is not valid

Calculation name:

4 Click **Calculate** to perform the calculation
or
Add more calculations to the **Calculation** list (see Chapter 6 for information about the workflow).

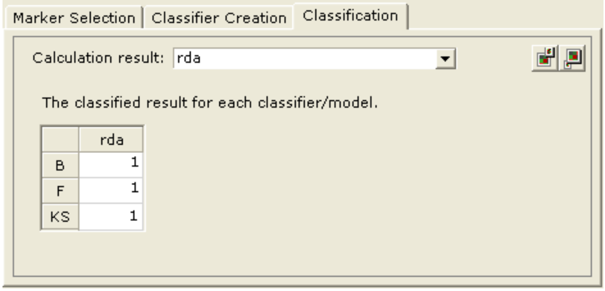
5 When the calculation(s) has/have finished, this will be indicated by a status icon in front of the calculation. The status of the calculations will also be displayed in the **Calculation status** field.

For information on how to analyze the results, see section 10.9.

10.9 Analyze the results of the Classification calculation

1. Select the **Results** step in the workflow area, select **Discriminant Analysis** in the Results bar and select the **Classification** tab in the Results view.

   The results are displayed in the Results view and in the Spot Map table.



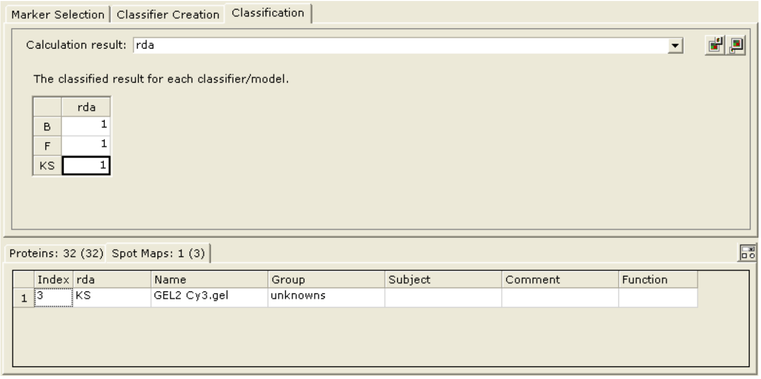
Calculation result: rda

The classified result for each classifier/model.

	rda
B	1
F	1
KS	1

   The number of spot maps classified to the different groups is displayed. The classification result (group name or condition name) for each spot map is shown in the calculation name column in the Spot Map table (the column with the same name as in the **Calculation result** field).

2. Click on a group to display only these spot maps in the Spot Map table.



Calculation result: rda

The classified result for each classifier/model.

	rda
B	1
F	1
KS	1

	Index	rda	Name	Group	Subject	Comment	Function
1	3	KS	GEL2 Cy3.gel	unknowns			

11 Interpretation

11.1 Overview

The fourth main step in the EDA analysis is **Interpretation**. In this step, biological information and context from in-house or public databases are integrated for the proteins of interest found in the results step. This step provides the possibility to check whether or not the results correspond to the biological findings. It can also reveal new hypotheses.

To be able to perform interpretation, protein ID including accession number (MS data) must be available for the proteins. If MS data exists for the proteins, these can be imported into the EDA workspace by combining the data with a pick list. MS data can also be entered manually in EDA.

If MS data was included in the BVA workspaces imported into EDA, this data was also imported into EDA.

Note: *The protein ID including accession number is denoted MS data in this manual.*

Note: *If no MS data exists for the proteins, it is possible to generate a pick list from a set in EDA and apply it to a pick gel in a corresponding BVA. The proteins can then be picked and MS analysis performed, generating MS data which can be imported into EDA.*

To interpret the results for a selected set of proteins, queries are created that perform specified searches in the appropriate databases. For example, it is possible to create a query that finds all pathways where the proteins are included by searching the available databases.

Depending on the results of the queries, new sets can be created to further reduce the data set.

If required, more calculations can be performed on the created sets in the **Calculations** step.

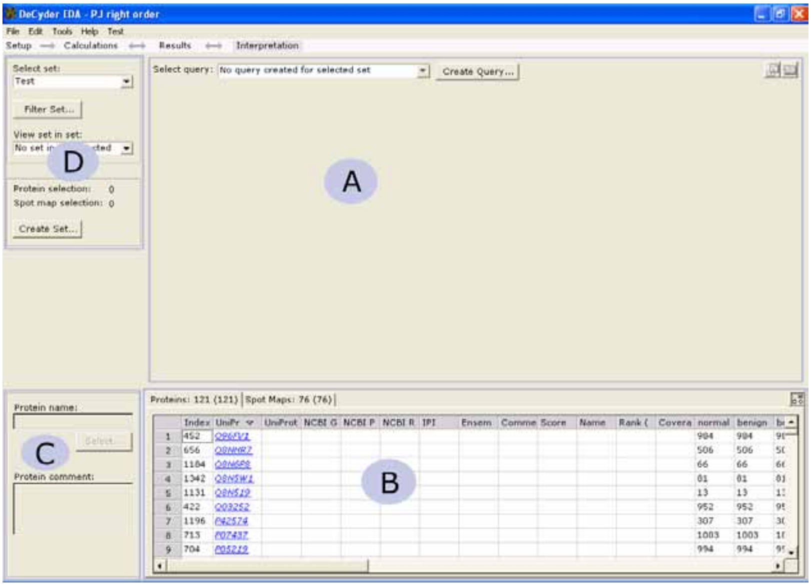
Creating queries, viewing the results of the queries and creating new sets are performed in the different areas of the **Interpretation** window:

- **Results view** (**A**)
  Create queries and view the results in this area.

- **Protein/Spot map table** (**B**)
  Shows information on proteins and spot maps in a table format.

- **Protein/Spot map details area** (**C**)
  Shows details on the protein/spot map selected in the protein/spot map table or results view.

- **Set area** (**D**)
  Select the set for which to set up queries and view results in the results view and protein/spot map table as well as creating new sets.


DeCyder EDA - PJ right order
File Edit Tools Help Test
Setup → Calculations ↔ Results ↔ Interpretation
Filter Set...
Create Set...
Create Query...
Select...
A
B
C
D

Select set: Test

View set in set: No set in ... cted

Protein selection: 0

Spot map selection: 0

Select query: No query created for selected set

Protein name:

Protein comment:

Proteins: 121 (121) | Spot Maps: 76 (76)

	Index	UniPr	UniProt	NCBI G	NCBI P	NCBI R	IPI	Ensem	Comme	Score	Name	Rank (	Covera	normal	benign	b
1	452	Q96FV1												984	984	9
2	656	Q8NNR7												506	506	5
3	1184	Q8N6P8												66	66	6
4	1342	Q8N5W1												01	01	01
5	1131	Q8N519												13	13	1
6	422	Q03252												952	952	9
7	1196	P42574												307	307	3
8	713	P07437												1003	1003	1
9	704	P05219												994	994	9

Fig 11-1. Interpretation window.

11.2 Workflow

Before performing interpretation, a set with proteins of interest with known ID and accession number must have been created.

An example of a workflow in the Interpretation step is outlined below:

1 **Create pick list, perform MS or MS/MS analysis and import MS data**
If no MS data exists for the proteins in the set of proteins on which to perform interpretation, generate a pick list, perform MS or MS/MS analysis and import the MS data. If MS data exists but has not been imported yet, import the MS data.

See section 11.3.

If MS data has already been imported into EDA proceed with step 2.

Note: *It is also possible to manually enter accession numbers for the different proteins in EDA, see section 11.3.3.*

2 **Perform Interpretation**
Create queries to get information from different databases on the protein and display this information in EDA.

See section 11.4.

3 **Use web links**
Click on the web links in the protein table to open the protein in the set database. It is possible to set which databases should be opened in the **Web Links Settings** dialog.

See section 11.5.

11.3 Create pick list and import MS data

11.3.1 Create pick lists

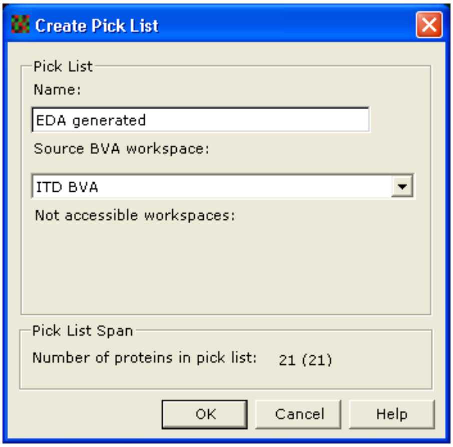
Pick lists can be generated for a set of proteins. If the proteins in a set come from different BVA workspaces, one pick list should be generated from the BVA where you have a pick gel you want to pick spots from for further preparation and MS analysis.

Note: *It is only necessary to generate a pick list if MS data is not available for the proteins on which to perform interpretation.*

To generate a pick list:

1 Select the set containing the proteins to be included in the pick list in the **Select set** field.

2 Select **Tools:Create Pick List in BVA ...** in the menu bar. The **Create Pick List** dialog is displayed.



Name: EDA generated

Source BVA workspace: ITD BVA

Not accessible workspaces:

Number of proteins in pick list: 21 (21)

3 Enter a **Name** for the pick list to be created.

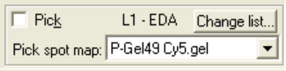
4 Select the BVA workspace that contains the proteins to pick from the drop-down list in the **Source BVA workspace** field. If a source BVA workspace is not available in the database, this will be displayed in the **Not accessible workspaces** field.

The **Pick List Span** area lists how many of the proteins in the set in EDA exist in the selected BVA workspace. For example, if 10 (10) is displayed, this means that 10 out of 10 proteins in the set in EDA exist in the selected Source BVA workspace.

Note: *If not all proteins in the set are included in the Pick List Span area (which may happen if the BVA workspaces are not linked or in some cases, if the workspaces are linked by Template) additional pick lists must be created, one/Source BVA workspace, to include the proteins in the set in EDA in at least one pick list.*

5 Click **OK** to create the pick list. BVA opens displaying the Source BVA workspace in the Protein Table view.

If a pick gel has already been set, it will be displayed together with the created pick list (otherwise set the appropriate spot map to Pick and define picking references, see *DeCyder 2D Software Version 6.5 User Manual*).

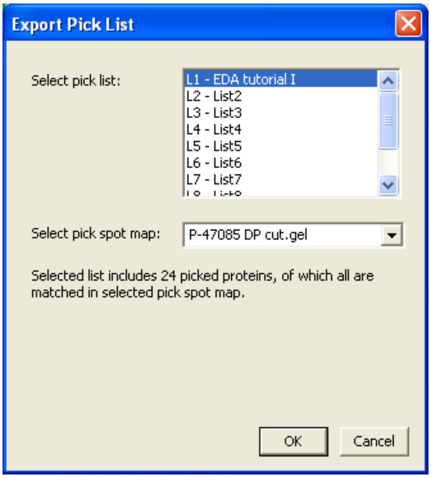


□ Pick L1 - EDA

Pick spot map: P-Gel49 Cy5.gel

Note: *Check that all spots in the pick list exported from EDA are matched on the pick gel in BVA and that picking references are detected. If not try to match all spots manually to the pick gel and add picking references before exporting the pick list. See DeCyder 2D Software Version 6.5 User Manual for more information.*

To export the pick list, select **File:Export Pick List...** in BVA.
The **Export Pick List** dialog opens with the correct pick list and pick spot map selected.



Export Pick List

Select pick list:
L1 - EDA tutorial I
L2 - List2
L3 - List3
L4 - List4
L5 - List5
L6 - List6
L7 - List7
L8 - List8

Select pick spot map: P-47085 DP cut.gel

Selected list includes 24 picked proteins, of which all are matched in selected pick spot map.

6 Click **OK**. In the dialog that appears, choose a folder in which to save the pick list, type a name for the pick list, make sure the file format is *.txt and click **Save**. The pick list has now been created.

Note: *In the case of linking by Template:*
If not all proteins in the set with proteins to pick were included in the created pick list, repeat steps 3-6 until the missing proteins can be found in another BVA.

Gel spots are then picked and prepared for MS/MS analysis or PMF analysis in a mass spectrometer. A protein search is then performed and the resulting MS data (in Sequest® or Mascot® format) can be imported into EDA.

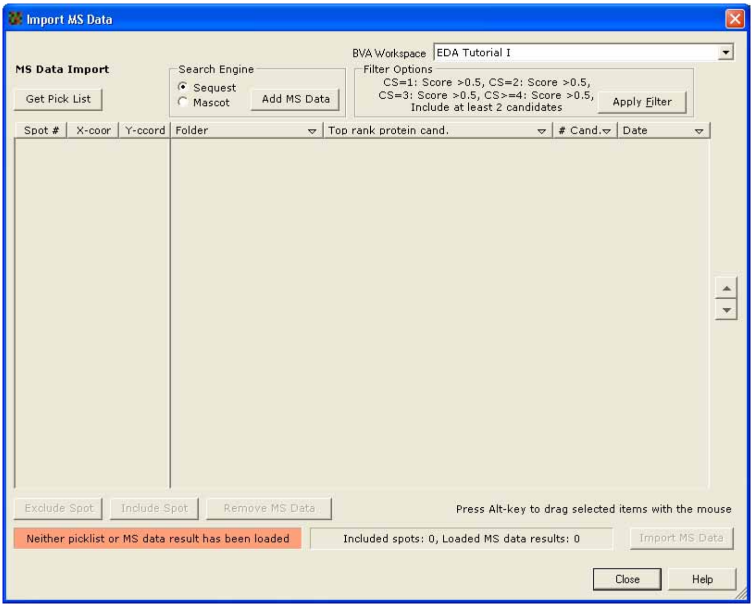
11.3.2 Import MS Data

Note: *If the EDA workspace already contains MS data for the proteins of interest, proceed with section 11.4.*

MS data (accession numbers) are needed to identify the proteins in the different databases used when performing interpretation.

To import MS data:

1 Select **File:Import MS data...** in the menu bar. The **Import MS Data** dialog is displayed.



BVA Workspace: EDA Tutorial I

Search Engine: ● Sequest ○ Mascot

Filter Options: CS=1: Score >0.5, CS=2: Score >0.5, CS=3: Score >0.5, CS>=4: Score >0.5, Include at least 2 candidates

Press Alt-key to drag selected items with the mouse

Neither picklist or MS data result has been loaded

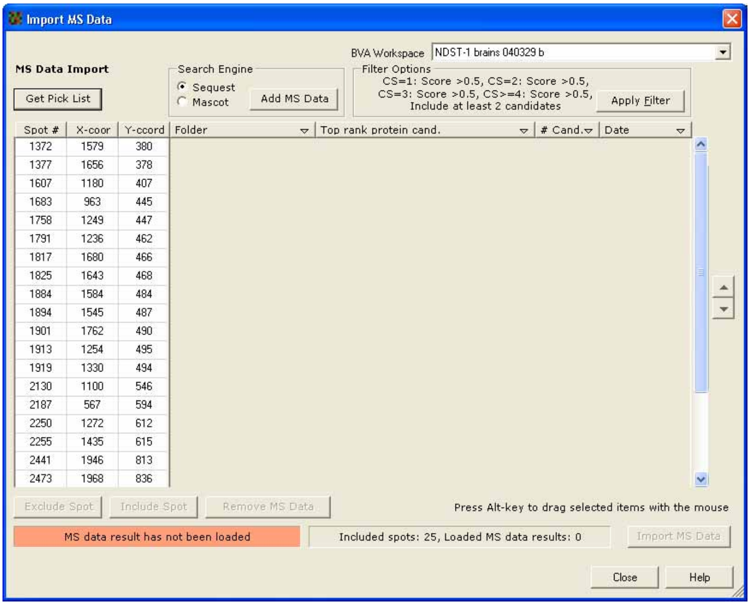
Included spots: 0, Loaded MS data results: 0

2 Select the BVA workspace in which the pick list was created in the **BVA workspace** drop-down list.

3 Click **Get Pick List** to open the **Get Pick List** dialog.

4 Locate the pick list (*.txt) on your computer from which to import MS data and click **OK** to import it.

The pick list is displayed in the left part of the dialog. The **Spot#** column shows the Master spot number of the protein and the **X-coord** and **Y-coord** columns show the coordinates on the pick gel.



BVA Workspace: NDST-1 brains 040329 b

Search Engine: ● Sequest ○ Mascot

Filter Options: CS=1: Score >0.5, CS=2: Score >0.5, CS=3: Score >0.5, CS>=4: Score >0.5, Include at least 2 candidates

Spot #	X-coor	Y-ccord	Folder	Top rank protein cand.	# Cand.	Date
1372	1579	380				
1377	1656	378				
1607	1180	407				
1683	963	445				
1758	1249	447				
1791	1236	462				
1817	1680	466				
1825	1643	468				
1884	1584	484				
1894	1545	487				
1901	1762	490				
1913	1254	495				
1919	1330	494				
2130	1100	546				
2187	567	594				
2250	1272	612				
2255	1435	615				
2441	1946	813				
2473	1968	836				

Press Alt-key to drag selected items with the mouse

MS data result has not been loaded

Included spots: 25, Loaded MS data results: 0

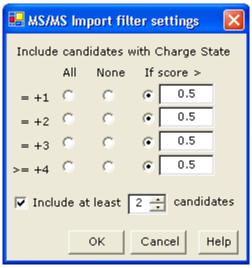
5 Select the type of MS data file to import by choosing the appropriate radio button in the **Search Engine** area.

Radio button	Description
Sequest	Select to import MS/MS data that has been received by searching in the Sequest database. The resultant file format is a MS/MS folder containing the *.out files from a search.
Mascot	Select to import MS data that has been received by searching in the Mascot database. The resultant file format is *.dat and was obtained either from a Peptide Mass Fingerprint search or an MS/MS search.

6 A default filter for the selected type of MS data is displayed in the **Filter Options** area.

7 If you want to change filter, click **Apply Filter** to select a filter for the MS data to import. Proceed with step 8. Otherwise, proceed with step 10.

Depending on the selected radio button in the **Search Engine** area, the **MS/MS Import filter settings** dialog or **MS Import filter settings** dialog is displayed.



MS/MS Import filter settings

Include candidates with Charge State

	All	None	If score >	
= +1	○	○	●	0.5
= +2	○	○	●	0.5
= +3	○	○	●	0.5
>= +4	○	○	●	0.5

☑ Include at least 2 candidates

8 If the **Sequest** radio button was selected, edit the settings in the **MS/MS Import filter settings** dialog as appropriate (default settings are shown in the screenshot):

a. Select which candidates for the different charge states to use by selecting the appropriate radio button for the charge state.

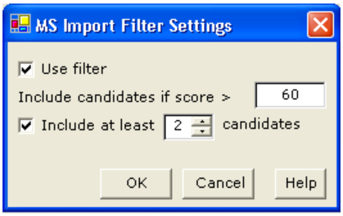
Radio button	Description
All	Select to include all protein candidates for the charge state.
None	Select to exclude any protein candidates for the charge state.
If score>	Select to include protein candidates with at least a certain score. Enter the appropriate score in the field to the right (or use the default values).

b. *To include at least a specified number of protein candidates regardless of score:*
Check the box **Include at least x candidates** and enter a number for how many candidates to include.

c. Click **OK** to set the filter and return to the **Import MS data** dialog. The filter is displayed in the **Filter Options** area.

9 If the **Mascot** radio button was selected, edit the settings as appropriate in the **MS Import Filter Settings** dialog (default settings are shown in the screenshot):

a. Check/uncheck the box **Use filter** as appropriate and if the box is checked, enter a score number in the **Include candidates if score >** field. Only protein candidates with a score higher than the number specified will be imported.



MS Import Filter Settings

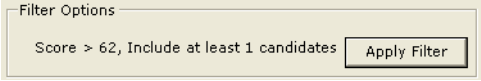
☑ Use filter

Include candidates if score > 60

☑ Include at least 2 candidates

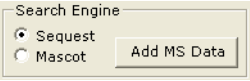
b. Check/uncheck the box **Include at least x candidates** as appropriate. If checked, enter how many candidates to include. If not enough candidates reach the specified score, this number of candidates will still be imported.

c. Click **OK** to set the filter and return to the **Import MS data** dialog. The filter is displayed in the **Filter Options** area.



Filter Options

Score > 62, Include at least 1 candidates



Search Engine

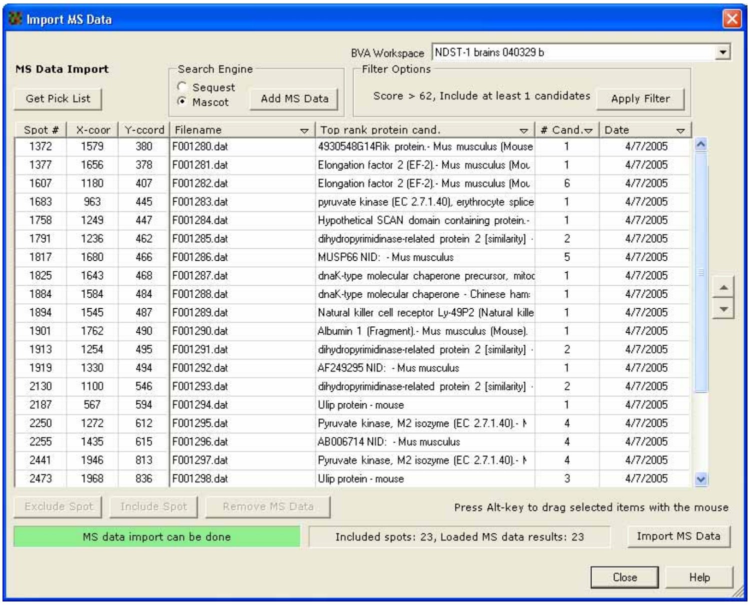
○ Sequest ● Mascot

10 Click **Add MS Data** in the **Search Engine** area. The **Select Sequest result files** dialog or **Select Mascot result files** dialog is displayed (depending on the selected radio button in the **Search Engine** area).

Sequest	In the Sequest case, folders containing .out files should be selected. 1 folder = 1 spot.
Mascot	In the Mascot case, .dat files should be selected. 1 file =1 spot.

11 Browse to locate and select the result files or folders to import (multiple files or folders can be selected using Ctrl + click or Shift + click) and click **OK**.

The data or MS/MS Data is displayed in the MS Data table below, next to the Pick List data.



Import MS Data

BVA Workspace: NDST-1 brains 040329 b

MS Data Import

Search Engine: ○ Sequest ● Mascot

Filter Options: Score > 62, Include at least 1 candidates

Spot #	X-coor	Y-ccord	Filename	Top rank protein cand.	# Cand.	Date
1372	1579	380	F001280.dat	4930548G14Rik protein.- Mus musculus (Mouse	1	4/7/2005
1377	1656	378	F001281.dat	Elongation factor 2 (EF-2).- Mus musculus (Mou	1	4/7/2005
1607	1180	407	F001282.dat	Elongation factor 2 (EF-2).- Mus musculus (Mou	6	4/7/2005
1683	963	445	F001283.dat	pyruvate kinase (EC 2.7.1.40), erythrocyte splice	1	4/7/2005
1758	1249	447	F001284.dat	Hypothetical SCAN domain containing protein.-	1	4/7/2005
1791	1236	462	F001285.dat	dihydropyrimidinase-related protein 2 [similarity] ·	2	4/7/2005
1817	1680	466	F001286.dat	MUSP66 NID: - Mus musculus	5	4/7/2005
1825	1643	468	F001287.dat	dnaK-type molecular chaperone precursor, mitoc	1	4/7/2005
1884	1584	484	F001288.dat	dnaK-type molecular chaperone - Chinese ham:	1	4/7/2005
1894	1545	487	F001289.dat	Natural killer cell receptor Ly-49P2 (Natural kille	1	4/7/2005
1901	1762	490	F001290.dat	Albumin 1 (Fragment).- Mus musculus (Mouse).	1	4/7/2005
1913	1254	495	F001291.dat	dihydropyrimidinase-related protein 2 [similarity] ·	2	4/7/2005
1919	1330	494	F001292.dat	AF249295 NID: - Mus musculus	1	4/7/2005
2130	1100	546	F001293.dat	dihydropyrimidinase-related protein 2 [similarity] ·	2	4/7/2005
2187	567	594	F001294.dat	Ulip protein - mouse	1	4/7/2005
2250	1272	612	F001295.dat	Pyruvate kinase, M2 isozyme (EC 2.7.1.40).- M	4	4/7/2005
2255	1435	615	F001296.dat	AB006714 NID: - Mus musculus	4	4/7/2005
2441	1946	813	F001297.dat	Pyruvate kinase, M2 isozyme (EC 2.7.1.40).- M	4	4/7/2005
2473	1968	836	F001298.dat	Ulip protein - mouse	3	4/7/2005

Press Alt-key to drag selected items with the mouse

MS data import can be done

Included spots: 23, Loaded MS data results: 23

The Filename/Folder name, the Top rank protein candidate name, the number of candidates that will be imported and the import date is displayed for the MS data. It is only possible to view all candidate proteins when the data has been imported into EDA.

12 *Match the MS data to the pick list if results were not imported in the same order as in the pick list:*
Select a row in the MS data table that is not matched and move it up or down in the list, using the arrows to the right, until it matches the correct spot in the Pick List table.

Note: *Preferably, the MS or MS/MS analysis of the picked spots have been run and/or imported in the same order as in the pick list, giving a correspondence between the MS data and the pick list master spot number. In this case, the matching of the rows in the two tables is already correct. However, the matching should always be checked.*

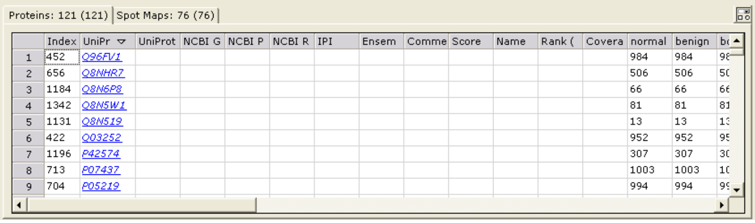
13 If MS data is not available for a spot in the pick list, exclude it by selecting the spot and clicking **Exclude**. The row in the Pick List table will turn gray indicating that the spot has been excluded and an empty gray row is inserted in the data table, shifting all the rows beneath one step down.

Tip! *To include an excluded spot again, select it and click Include. The row in the Pick List table will turn black, indicating that the spot is included and the empty gray row will disappear in the data table, shifting all the rows beneath one step up.*

14 If you need to remove a MS data result (row), select it and click **Remove MS data**.

15 Click **Import MS data** to import the matched MS data. This button is enabled only when all active (black) rows in the pick list have a corresponding file to the right.

The Protein Table in EDA will be updated with the MS data information.



Proteins: 121 (121) | Spot Maps: 76 (76)

	Index	UniPr	UniProt	NCBI G	NCBI P	NCBI R	IPI	Ensem	Comme	Score	Name	Rank (	Covera	normal	benign	bc
1	452	Q96FV1												984	984	98
2	656	Q8NHR7												506	506	50
3	1184	Q8N6P8												66	66	66
4	1342	Q8N5W1												81	81	81
5	1131	Q8N519												13	13	13
6	422	Q03252												952	952	95
7	1196	P42574												307	307	30
8	713	P07437												1003	1003	10
9	704	P05219												994	994	99

The accession numbers for the proteins will be filled in and web links will be displayed. It is possible to click on the web links and open the proteins in the specified databases. It is also possible to right-click on a web link and select database in which to open the protein. See section 11.5 for more information on web links.

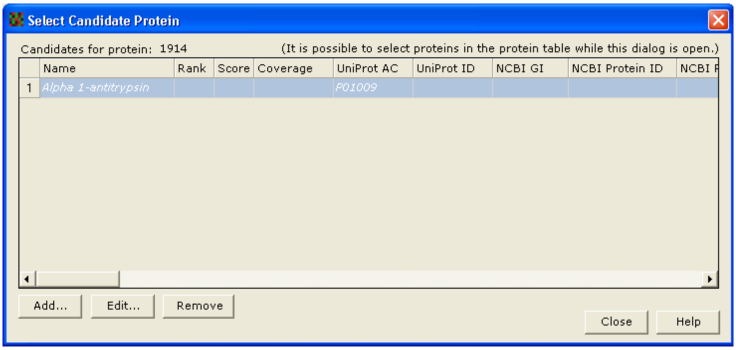
11.3.3 Enter/edit candidate proteins in EDA

In EDA it is possible to change which of the imported candidate proteins to use as well as add/edit candidate proteins. For example, if MS data has not been imported it is possible to add candidate proteins manually. All of this can be performed in the **Select Candidate Protein** dialog.

1 Select the protein for which to select or add/edit candidate proteins in the Protein table.

2 *To add candidate proteins manually if no MS data has been imported:*

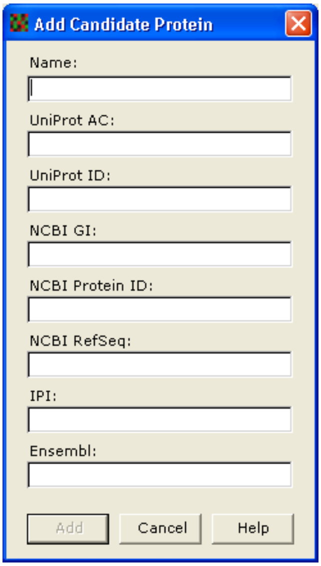
a. Click the **Select...** button in the Protein details area to the left. The **Select Candidate Protein** dialog is displayed.



Candidates for protein: 1914 (It is possible to select proteins in the protein table while this dialog is open.)

	Name	Rank	Score	Coverage	UniProt AC	UniProt ID	NCBI GI	NCBI Protein ID	NCBI F
1	Alpha 1-antitrypsin				P01009				

b. Click **Add...** . The **Add Candidate Protein** dialog is displayed.



c. Enter a name for the protein.

d. Fill in the appropriate accession numbers in their corresponding fields.

   At least one of the accession number type fields must be filled in.

e. Click **Add** to add the protein candidate and return to the **Select Candidate Protein** dialog. The candidate protein will be displayed in the **Select Candidate Protein** dialog.

   A manually added candidate protein will be displayed in italic style.

f. Click **Close** in the **Select Candidate Protein** dialog to apply the settings.

3 *To select another candidate protein than the top-ranked or to add/edit protein candidates imported into EDA:*

a. Click the **Select...** button in the Protein details area to the left. The **Select Candidate Protein** dialog is displayed.

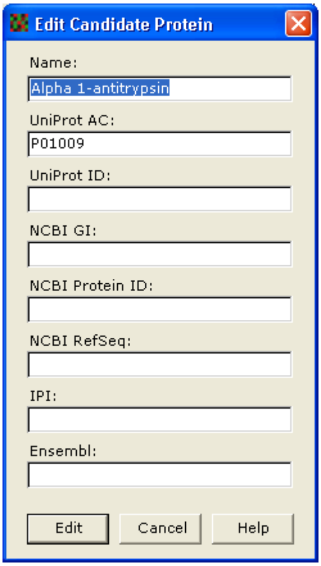
b. *To select another candidate protein other than the top-ranked among the imported candidate proteins*, click on the appropriate one in the list and click **Close**.

c. *To add a new candidate protein*, click **Add...** . The **Add Candidate Protein** dialog is displayed.
   Enter a name for the protein, fill in the appropriate accession numbers and click **Add**.

d. *To edit a candidate protein*, select the candidate protein in the **Select Candidate Protein** dialog and click **Edit**.

The **Edit Candidate Protein** dialog is displayed.

Edit the name and/or accession numbers as appropriate and click **Edit**.


Edit Candidate Protein

Name: Alpha 1-antitrypsin

UniProt AC: P01009

UniProt ID:

NCBI GI:

NCBI Protein ID:

NCBI RefSeq:

IPI:

Ensembl:

e. *To remove a candidate protein*, select the candidate protein in the **Select Candidate Protein** dialog and click **Remove**.

f. Click **Close** in the **Select Candidate Protein** dialog to apply the settings.

Note: *A manually added/edited candidate protein will be displayed in italic style.*

11.4 Perform interpretation

Interpretation is a very powerful tool used to get information from different databases about the proteins of interest and present the information in a user-friendly way in EDA.

When performing interpretation, queries are created that are designed to extract information from specific databases depending on the available accession numbers. The result of the queries are displayed in the Interpretation window.

Note: *If your system uses a proxy server for internet access, the correct proxy settings must be entered in order to create queries. Furthermore, in order to create PubMed queries, the discoveryHub software needs to be installed. See Appendix G for more information on how to enter settings for proxy access and discoveryHub.*

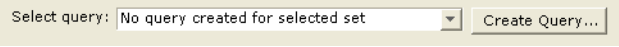
11.4.1 Create queries

Create queries to obtain information on, for example, the molecular functions of the protein, biological processes, which pathways the proteins are part of and so on.

To create a query:

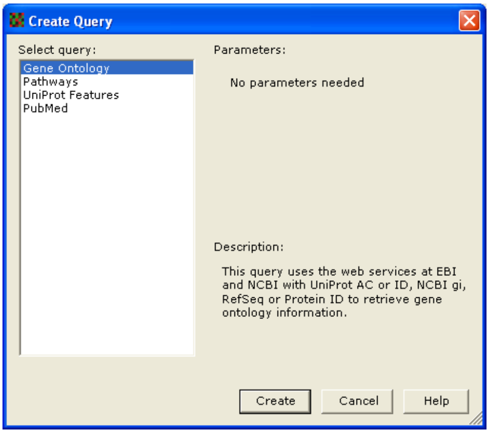
1. In the **Select set** drop-down list, select the set which contains the proteins for which you want to create queries (e.g. markers found in the discriminant analysis).

2. In the **Select query** field, click **Create query** to create a new query.



Select query: No query created for selected set

3. The **Create query** dialog opens.



No parameters needed

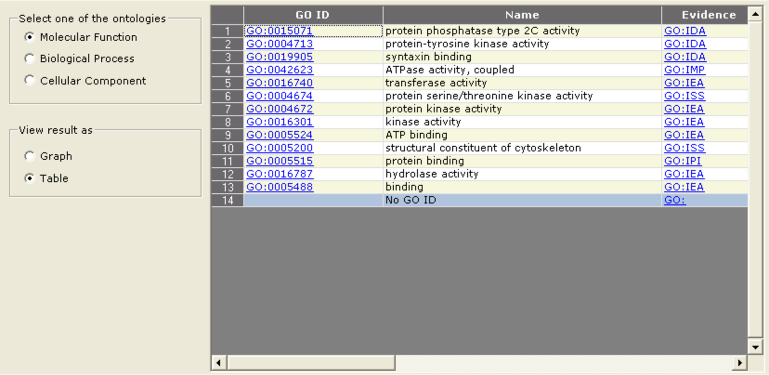
This query uses the web services at EBI and NCBI with UniProt AC or ID, NCBI gi, RefSeq or Protein ID to retrieve gene ontology information.

Four queries are available by default: **Gene Ontology**, **Pathways**, **UniProt Feature** and **PubMed.**

Query	**Description**
Gene Ontology	Select this query to get information from the database (provided by the Gene Ontology Consortium) on each protein's: • molecular functions (for example catalytical activity, transporter activity, binding etc.) • if the protein is involved in one or more biological process (for example cell growth and maintenance, signal transduction etc.) • if the protein is part of any cellular component (the nucleus, for example)
Pathways	Select this query to get information from the KEGG database on which pathways the proteins are part of.
UniProt Feature	Select this query to get information on the proteins from the UniProt Features database.
PubMed	Select this query to get articles from PubMed on the different proteins. ***Note:*** *To be able to use this query, a license for discoveryHub must be purchased from GE Healthcare. Information on installing the software will be provided with the software.* *However, settings for discoveryHub must be set in the Database Administration Tool in DeCyder 2D Software. See Appendix G, DiscoveryHub for more information.*

Note: *It is possible to design your own queries. Please contact GE Healthcare for more information.*

4 Select the appropriate query and click **Create** to add it to the **Select query** list in the **Interpretation** window. The results of the query for the selected proteins are displayed below the list. Fig. 11-2 shows an example of the results of a **Gene Ontology** query.



Select one of the ontologies

● Molecular Function

○ Biological Process

○ Cellular Component

View result as

○ Graph

● Table

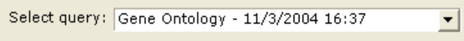
	GO ID	Name	Evidence
1	GO:0015071	protein phosphatase type 2C activity	GO:IDA
2	GO:0004713	protein-tyrosine kinase activity	GO:IDA
3	GO:0019905	syntaxin binding	GO:IDA
4	GO:0042623	ATPase activity, coupled	GO:IMP
5	GO:0016740	transferase activity	GO:IEA
6	GO:0004674	protein serine/threonine kinase activity	GO:ISS
7	GO:0004672	protein kinase activity	GO:IEA
8	GO:0016301	kinase activity	GO:IEA
9	GO:0005524	ATP binding	GO:IEA
10	GO:0005200	structural constituent of cytoskeleton	GO:ISS
11	GO:0005515	protein binding	GO:IPI
12	GO:0016787	hydrolase activity	GO:IEA
13	GO:0005488	binding	GO:IEA
14		No GO ID	GO:

Fig 11-2. Example of the results from the Gene Ontology query.

5 To create more queries, repeat from step 2 and select a new query. All created queries are placed in the **Select query** pop-up dialog and it is possible to view the results for the queries by choosing a query from the list. Proceed with section 11.4.2 for information on the results for the different queries.

11.4.2 View query results

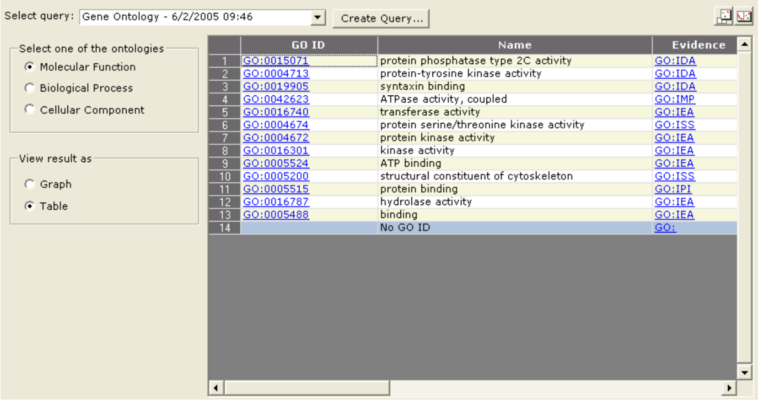
All queries that have been created will be available in the **Select query** pop-up dialog. The creation date and time is shown for the query.



Select query: Gene Ontology - 11/3/2004 16:37

11.4.3 View results for Gene Ontology

1 Select the Gene Ontology query in the **Select query** pop-up dialog. The results for the query are displayed in the Gene Ontology results view.



Select query: Gene Ontology - 6/2/2005 09:46

Select one of the ontologies
- ● Molecular Function
- ○ Biological Process
- ○ Cellular Component

View result as
- ○ Graph
- ● Table

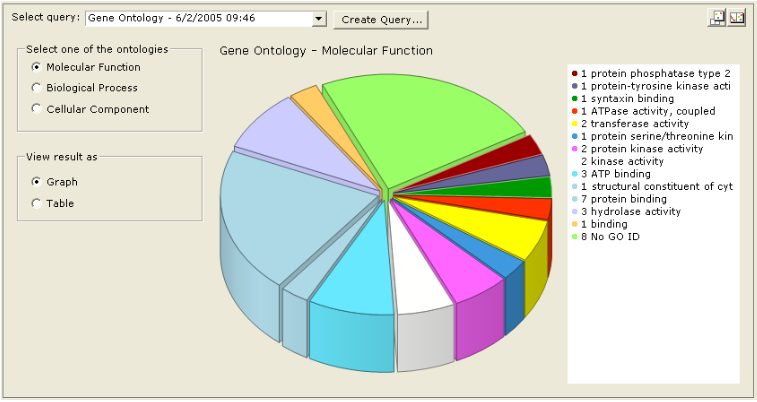
	GO ID	Name	Evidence
1	GO:0015071	protein phosphatase type 2C activity	GO:IDA
2	GO:0004713	protein-tyrosine kinase activity	GO:IDA
3	GO:0019905	syntaxin binding	GO:IDA
4	GO:0042623	ATPase activity, coupled	GO:IMP
5	GO:0016740	transferase activity	GO:IEA
6	GO:0004674	protein serine/threonine kinase activity	GO:ISS
7	GO:0004672	protein kinase activity	GO:IEA
8	GO:0016301	kinase activity	GO:IEA
9	GO:0005524	ATP binding	GO:IEA
10	GO:0005200	structural constituent of cytoskeleton	GO:ISS
11	GO:0005515	protein binding	GO:IPI
12	GO:0016787	hydrolase activity	GO:IEA
13	GO:0005488	binding	GO:IEA
14		No GO ID	GO:

By default, the results of the **Molecular Function** ontology are displayed in a table format (the radio button **Molecular Function** is selected in the **Select one of the ontologies** area and the **Table** radio button is selected in the **View results as** area).

2 To view the results of another ontology, select the appropriate radio button in the **Select one of the ontologies** area. The results for the ontology will be displayed in the results view. The table below summarizes the information displayed for the different ontologies.

Column	Description
GO ID	Shows the gene ontology ID of the protein. Click the link to open the protein in the database.
Name	Shows a description of the protein's molecular function (e.g. DNA binding protein), biological processes (e.g. DNA methylation) or cellular component (e.g. nucleus) depending on the selected ontology.
Evidence	Click the link to open the database and get information on the evidence for the protein, for example publications/articles.
Number of proteins	Shows how many proteins in the set that have this function. If clicking the row, the protein(s) in the protein table with this function, biological process or cellular component will be selected.

3 It is possible to display the results of the ontologies in a graph format by selecting **Graph** in the **View results as** area.



Select query: Gene Ontology - 6/2/2005 09:46

Select one of the ontologies
- ● Molecular Function
- ○ Biological Process
- ○ Cellular Component

View result as
- ● Graph
- ○ Table

Create Query...
Gene Ontology - Molecular Function
● 1 protein phosphatase type 2
● 1 protein-tyrosine kinase acti
● 1 syntaxin binding
● 1 ATPase activity, coupled
● 2 transferase activity
● 1 protein serine/threonine kin
● 2 protein kinase activity
2 kinase activity
● 3 ATP binding
● 1 structural constituent of cyt
● 7 protein binding
● 3 hydrolase activity
● 1 binding
● 8 No GO ID

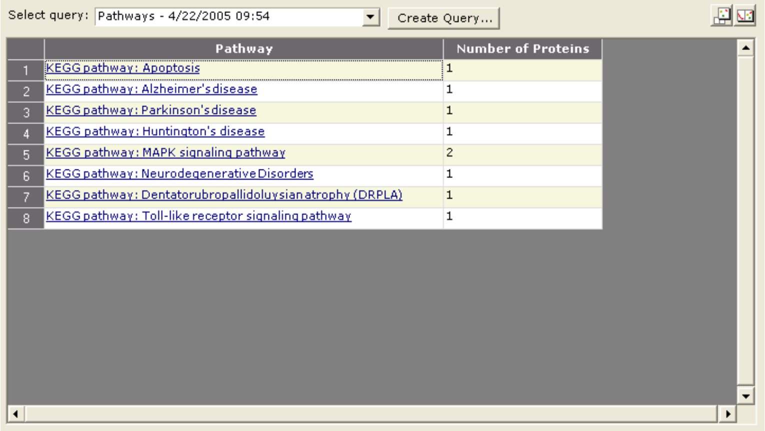
Then, the number of proteins:

- with certain molecular functions
- involved in certain biological processes
- that are part of a certain cellular components

are displayed in a graphical format (depending on the selected ontology). If performing a query on a large number of proteins, this display gives a nice overview of the different categories of proteins for the different ontologies.

11.4.4 View results for Pathways

1 Select the **Pathways** query in the **Select query** pop-up dialog. The results for the query are displayed in the Pathways results view.

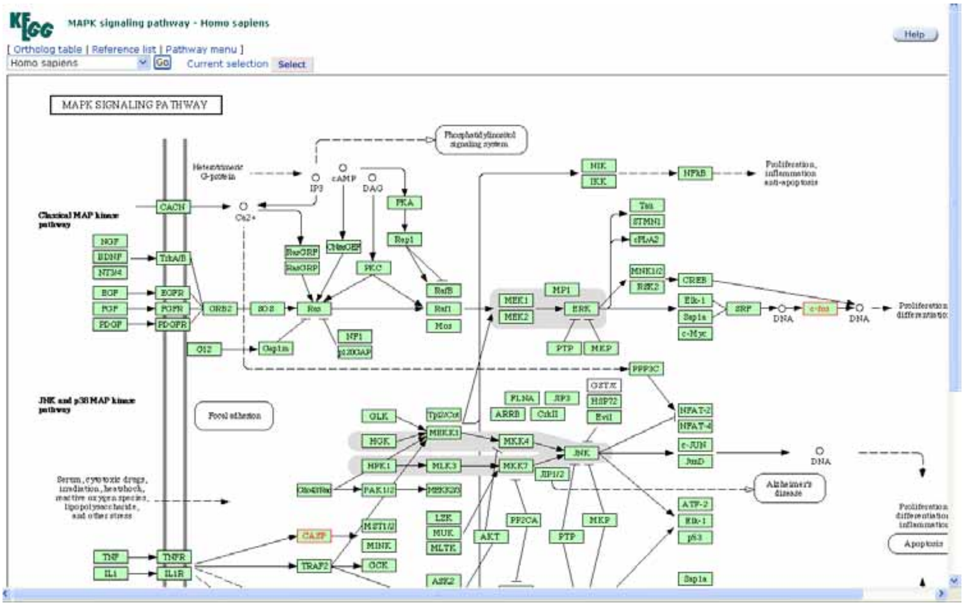


Select query: Pathways - 4/22/2005 09:54

	Pathway	Number of Proteins
1	KEGG pathway: Apoptosis	1
2	KEGG pathway: Alzheimer's disease	1
3	KEGG pathway: Parkinson's disease	1
4	KEGG pathway: Huntington's disease	1
5	KEGG pathway: MAPK signaling pathway	2
6	KEGG pathway: Neurodegenerative Disorders	1
7	KEGG pathway: Dentatorubropallidoluysian atrophy (DRPLA)	1
8	KEGG pathway: Toll-like receptor signaling pathway	1

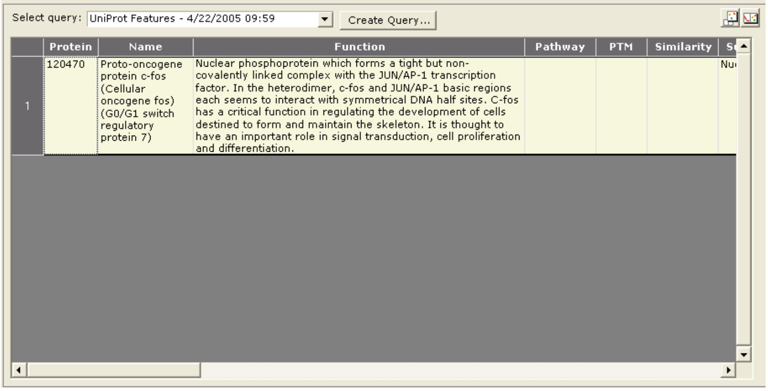
2 Clicking a row in the **Pathways** table will display the proteins in the Number of proteins row in the Protein table below. Clicking on a protein in the Protein table will highlight the protein in the **Pathways** table.

The **Pathway** column shows the different pathways found by the query and the number of proteins involved in the pathway. Clicking on the link will open the pathway with the protein(s) marked in red.


MAPK signaling pathway - Homo sapiens
[Ortholog table | Reference list | Pathway menu]
Homo sapiens
Go
Current selection
Select
Help
MAPK SIGNALING PATHWAY
Phosphatidylinositol signaling system
Heterotrimeric G-protein
IP3
cAMP
DAG
Ca2+
Classical MAP kinase pathway
CACN
PKA
Rap1
NGF
BDNF
NT3/4
TrkA/B
RasGRF
RasGRP
CNrasGEF
PKC
EGF
FGF
PDGF
EGFR
FGFR
PDGFR
GRB2
SOS
Ras
RafB
Raf1
Mos
NF1
p120GAP
G12
Gap1m
MEK1
MEK2
MP1
ERK
PTP
MKP
NIK
IKK
NFkB
Proliferation, inflammation anti-apoptosis
Tau
STMN1
cPLA2
MNK1/2
RSK2
CREB
Elk-1
Sap1a
c-Myc
SRF
DNA
c-fos
DNA
Proliferation differentiation
PPP3C
JNK and p38 MAP kinase pathway
Focal adhesion
GLK
HGK
HPK1
Tpl2/Cot
MEKK1
MLK3
GSTπ
HSP72
Evil
FLNA
JIP3
ARRB
CrkII
MKK4
MKK7
JNK
JIP1/2
NFAT-2
NFAT-4
c-JUN
JunD
DNA
Alzheimer's disease
Serum, cytotoxic drugs, irradiation, heatshock, reactive oxygen species, lipopolysaccharide, and other stress
Cdc42/Rac
PAK1/2
MEKK2/3
MST1/2
MINK
GCK
LZK
MUK
MLTK
ASK2
AKT
PP2CA
PTP
MKP
ATF-2
Elk-1
p53
Sap1a
Proliferation differentiation inflammation
Apoptosis
CASP
TNF
IL1
TNFR
IL1R
TRAF2

11.4.5 View results for UniProt Features

1 Select the appropriate **UniProt Features** query in the **Select query** pop-up dialog. The results for the query are displayed in the UniProt Features results view.



Select query: UniProt Features - 4/22/2005 09:59

	Protein	Name	Function	Pathway	PTM	Similarity	Su
1	120470	Proto-oncogene protein c-fos (Cellular oncogene fos) (G0/G1 switch regulatory protein 7)	Nuclear phosphoprotein which forms a tight but non-covalently linked complex with the JUN/AP-1 transcription factor. In the heterodimer, c-fos and JUN/AP-1 basic regions each seems to interact with symmetrical DNA half sites. C-fos has a critical function in regulating the development of cells destined to form and maintain the skeleton. It is thought to have an important role in signal transduction, cell proliferation and differentiation.				Nu

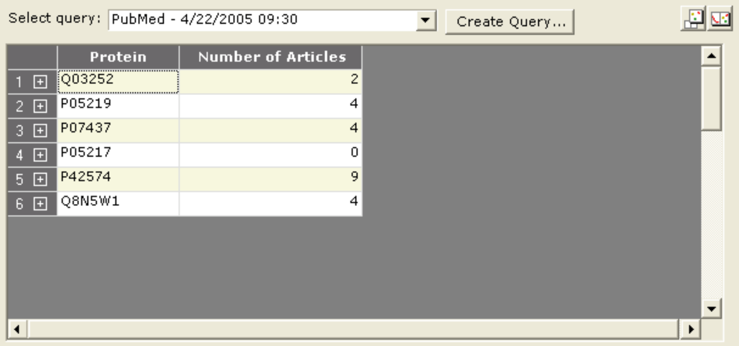
2 Click on a row in the table to show only that protein in the Protein table. Click on a protein in the Protein table to show the UniProt Features for that protein in the UniProt Features table.

3 Click the link in the Protein table to open the protein in the UniProt database and get detailed information for the protein and cross-references to other databases. A description of the columns are listed below.

Command	Description
Protein	Shows the protein accession number.
Name	Shows the name of the protein.
Function	Shows a general description of the function(s) of the protein.
Pathway	Shows a description of the metabolic pathways with which the protein is associated (if any).
PTM	Shows a description of a post-translational modification (if any).
Similarity	Shows a description of the similaritie(s) (sequence or structural) of the protein with other proteins (if any).
Sub-cellular location	Shows a description of the sub-cellular location of the mature protein (if available).
Keywords	Shows keywords that describe the protein.

11.4.6 View results from PubMed

1 Select the appropriate **PubMed** query in the **Select query** pop-up dialog. The results for the query are displayed in the PubMed results view.



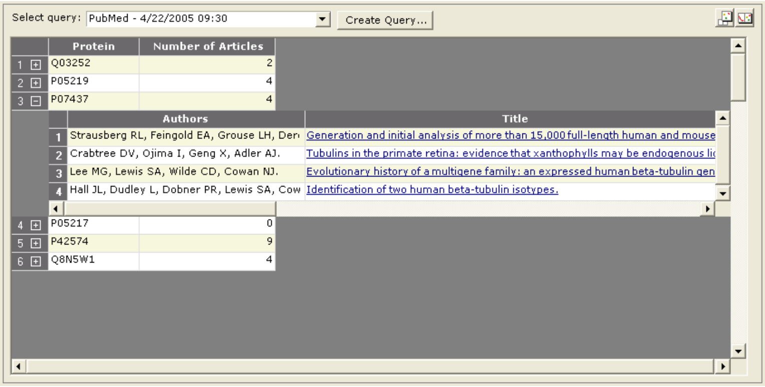
Select query: PubMed - 4/22/2005 09:30

	Protein	Number of Articles
1 ⊞	Q03252	2
2 ⊞	P05219	4
3 ⊞	P07437	4
4 ⊞	P05217	0
5 ⊞	P42574	9
6 ⊞	Q8N5W1	4

The results show the protein accession numbers and the number of publications/articles for each protein.

Selecting a row in the PubMed table will display this protein in the Protein table. Clicking a protein(s) in the Protein table will highlight the protein(s) in the PubMed table.

2 Click on the plus sign to expand the list and show the Authors, title and journal for each article.



Select query: PubMed - 4/22/2005 09:30

	Protein	Number of Articles
1 ⊞	Q03252	2
2 ⊞	P05219	4
3 ⊟	P07437	4

	Authors	Title
1	Strausberg RL, Feingold EA, Grouse LH, Derı	Generation and initial analysis of more than 15,000 full-length human and mouse
2	Crabtree DV, Ojima I, Geng X, Adler AJ.	Tubulins in the primate retina: evidence that xanthophylls may be endogenous li
3	Lee MG, Lewis SA, Wilde CD, Cowan NJ.	Evolutionary history of a multigene family: an expressed human beta-tubulin gen
4	Hall JL, Dudley L, Dobner PR, Lewis SA, Cow	Identification of two human beta-tubulin isotypes.

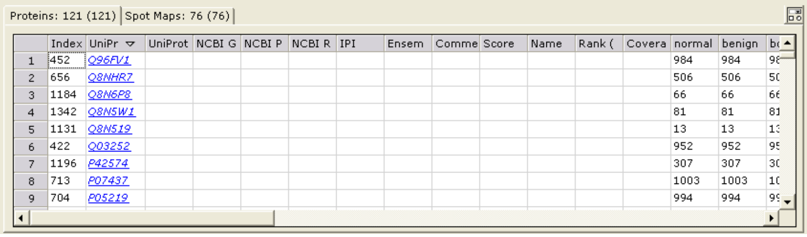
	Protein	Number of Articles
4 ⊞	P05217	0
5 ⊞	P42574	9
6 ⊞	Q8N5W1	4

3 Click the **Title** links to open the articled in PubMed and click the **Journal** links to open the journal's web site.

11.5 Using Web Links

When MS data has been imported, the protein accession numbers will be displayed in the form of web links in their respective columns in the Protein table.

The following types of accession numbers are supported in EDA: Uniprot AC, Uniprot ID, NCBI GI, NCBI Protein ID, NCBI RefSeq, Ensembl and IPI.



	Index	UniPr	UniProt	NCBI G	NCBI P	NCBI R	IPI	Ensem	Comme	Score	Name	Rank (	Covera	normal	benign	b
1	452	Q96FV1												984	984	98
2	656	Q8NHR7												506	506	50
3	1184	Q8N6P8												66	66	66
4	1342	Q8N5W1												81	81	81
5	1131	Q8N519												13	13	13
6	422	Q03252												952	952	95
7	1196	P42574												307	307	30
8	713	P07437												1003	1003	10
9	704	P05219												994	994	99

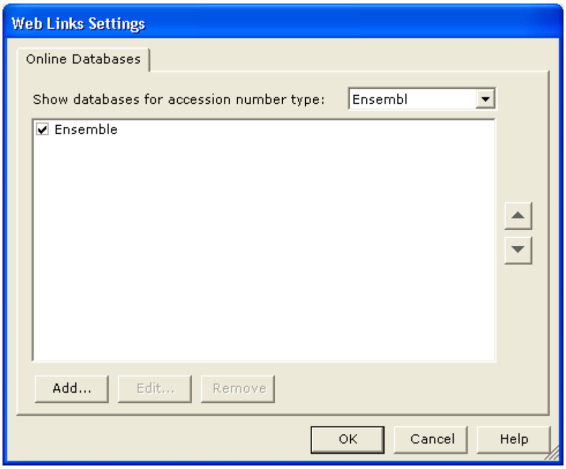
Clicking a web link will open the protein in the specified database. If several databases are available for an accession number, right-click the web link and select the appropriate database from which to open the protein.

Default databases are set for the different accession number types.

To add/edit databases to use for the accession number types, see section 11.5.1, Edit web links settings.

11.5.1 Edit web links settings

1 Select **Tools:Manage Web Links...** in the menu bar. The **Web Links Settings** dialog opens.

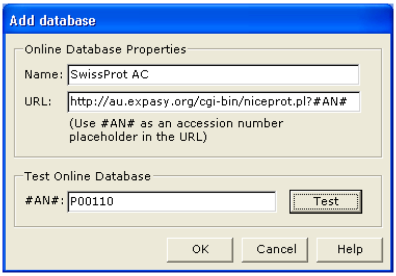


Show databases for accession number type: Ensembl

☑ Ensemble

2 Select the accession number type in the **Show databases for accession number type** drop-down list in order to display the available databases in the field below.

3 *To add a database for the selected accession number type:*



Name: SwissProt AC

URL: http://au.expasy.org/cgi-bin/niceprot.pl?#AN#

(Use #AN# as an accession number placeholder in the URL)

#AN#: P00110

a. Click **Add...** . The **Add database** dialog opens.

b. Enter a name for the database in the **Name** field.

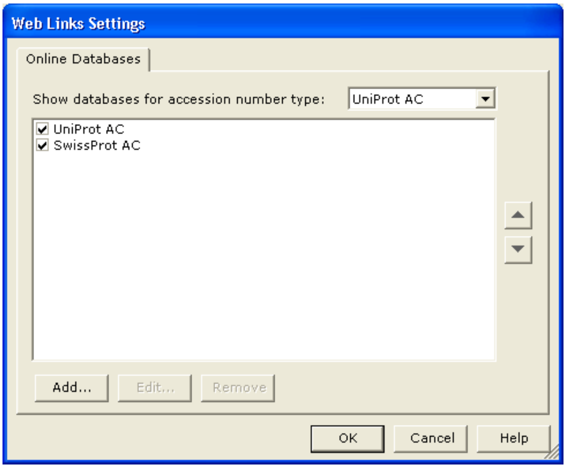
c. Enter the URL for the database.

Tip! *To find out what URL to enter, visit the official web page of the database using your internet browser, enter the accession number and perform the search. When the results are displayed, the URL is displayed in the browser. Copy this URL, paste it into the URL field in EDA and exchange the accession number in the URL to #AN#.*

d. The accession number type selected in the **Web Links Settings** dialog is displayed in the **Accession number type** field. Use this setting.

e. To test that the added database is correctly entered, enter an accession number for a protein in the **#AN#** field and click **Test**. The protein should be opened in the set database.

f. Click **OK** to add the database. It will appear in the **Web Links Settings** dialog.



Show databases for accession number type: UniProt AC

☑ UniProt AC

☑ SwissProt AC

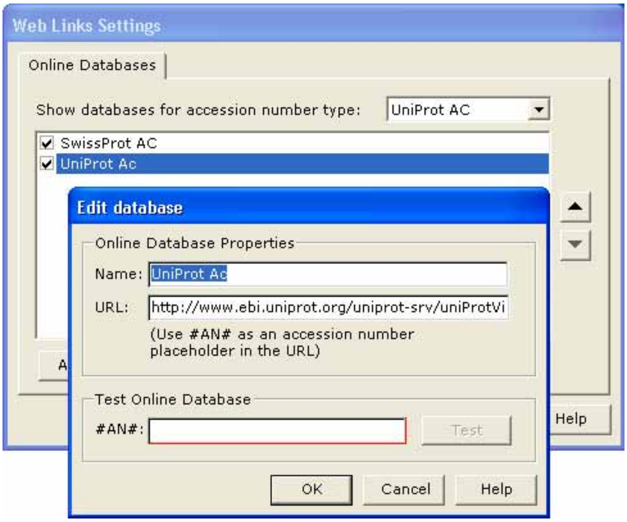
4 *To select which databases to use and change the order of databases:*

a. Check the boxes in front of the databases to be used to include them in the web links. Clicking a web link will open the protein in the first database in the list by default.

b. To change the order of databases, select a database and use the arrows to the right to move the database up or down in the list.

5 *To edit a database:*

a. Select the database to edit and click **Edit...** . The **Edit database** dialog opens.


Web Links Settings
Online Databases
Show databases for accession number type:
UniProt AC
☑ SwissProt AC
☑ UniProt Ac
Help
Edit database
Online Database Properties
Name:
UniProt Ac
URL:
http://www.ebi.uniprot.org/uniprot-srv/uniProtVi
(Use #AN# as an accession number placeholder in the URL)
Test Online Database
#AN#:
Test
OK
Cancel
Help

b. Edit the settings as required and click **OK**.

12 Creating and managing sets

12.1 Creating sets

New sets can be created in the **Calculation**, **Results** or **Interpretation** steps.

Based on the results of the calculations/interpretation, data can be selected directly (only in the **Results** and **Interpretation** steps) and a new set can be created including this data, or the data can be filtered and a new set with the filtered data can be created. As many sets as required can be created.

Sets can also be combined by using the logical conditions **AND** and **OR** to create a new set (see section 12.2, Managing sets).

12.1.1 Create a set by selecting data

Create a set by *selecting* data can only be performed in the **Results** and **Interpretation** steps (not in the **Calculation** step).

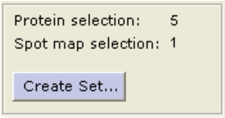
1 In the **Results**/**Interpretation** window, select the data to be included or excluded from the set to be created by:

- clicking directly on the proteins/spot maps in the heat map/graph

  and/or

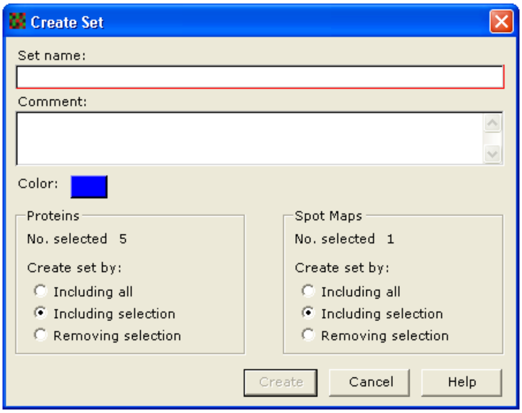
- selecting the proteins/spot maps by clicking them in the tables

Several proteins/spot maps can be selected by holding down the **Ctrl** or **Shift** key and clicking them in the graphs/tables.



Protein selection: 5
Spot map selection: 1

2 Click **Create Set...** in the left panel of the **Results/Interpretation** window. The **Create set** dialog opens.



Create Set

Set name:

Comment:

Color: 🟦

Proteins
No. selected 5
Create set by:
○ Including all
● Including selection
○ Removing selection

Spot Maps
No. selected 1
Create set by:
○ Including all
● Including selection
○ Removing selection

3 Enter a name for the set. If required, enter a comment on the set and select a color representing the set by clicking the color button and choosing the appropriate color.

Tip! *Different colors for the sets facilitates the analysis of the results in the results step.*

4 In the **Proteins** area, select the protein selection to include or exclude when creating a new set or to include all proteins in the new set to be created by choosing the appropriate radio button.

Note: *If proteins have been selected, the* ***Including selection*** *radio button is selected by default, otherwise* ***Including all*** *is selected by default*.

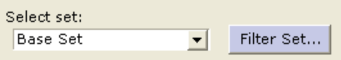
5 In the **Spot Maps** area, select the spot map selection to include or exclude when creating a new set or include all spot maps in the new set to be created by choosing the appropriate radio button.

Note: *If spot maps have been selected, the* ***Including selection*** *radio button is selected by default, otherwise* ***Including all*** *is selected by default*.

6 Click **Create** to create the set. The set will be added to the workspace but the previous set will still be displayed in the **Results/Interpretation** window.

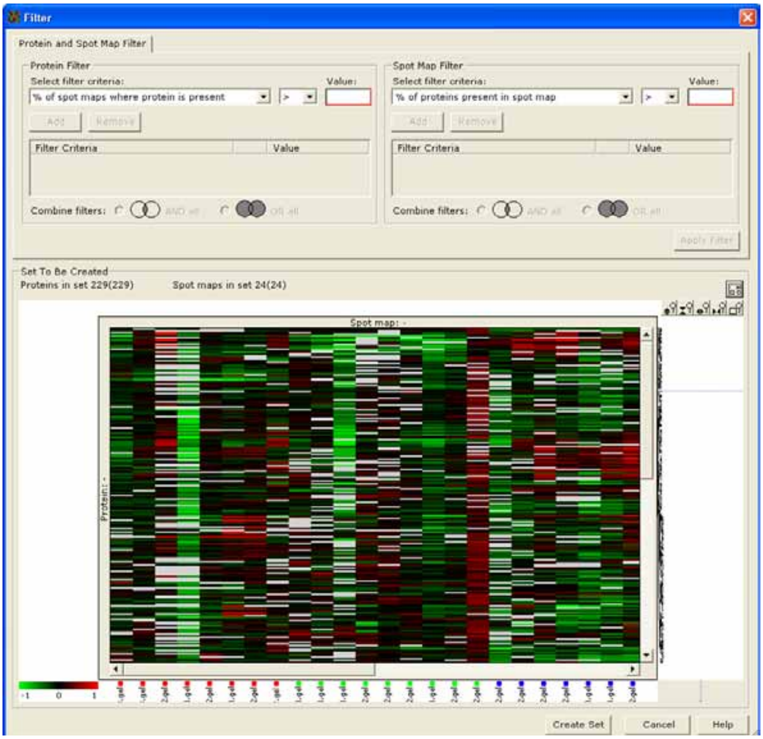
12.1.2 Create a set by filtering data

1 In the left panel of the **Calculations/Results/Interpretation** window, click **Filter Set...**.



Select set: Base Set

The **Filter** dialog opens.



2 *Define the appropriate protein filter criteria:*

a. Select filter criteria and **<**, **<=**, **=**, **>=** or **>** from the drop-down lists in the **Select filter criteria** field. For information about the different criteria, see section 12.1.3, Protein filter criteria.

b. Enter a value for the criteria in the **Value** field and click **Add** to add the filter criteria to the list. If required, repeat steps 2a and 2b to add a new protein filter criteria to the list.

c. Combine the **Protein Filter** criteria for the protein filter by using the logical conditions **AND all** or **OR all**.

Note: *Once a differential expression analysis calculation (for example ANOVA) has been performed, it can be used as a filter to extract proteins based on p-value. It will appear in the* ***Select filter criteria*** *drop-down list.*

3 *Define the appropriate spot map filter criteria:*

a. Select filter criteria and **<**, **<=**, **=**, **>=** or **>** (only for **% of proteins present in spot map**) from the drop-down lists in the **Select filter criteria** field. For information about the different criteria, see section 12.1.4, Spot Map filter criteria.

b. Enter a value for the criteria in the **Value** field and click **Add** to add the filter criteria to the list. If required, repeat steps 3a and 3b to add a new spot map filter criteria to the list.

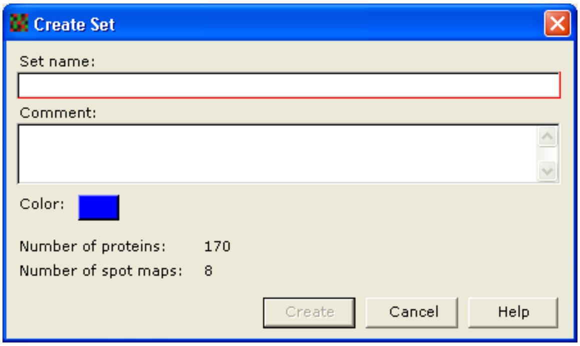
c. Combine the filter criteria for the spot map filter by using the logical conditions **AND all** or **OR all**.

Radio button	Description
AND all	Includes only those proteins that have been extracted by all filter criteria.
OR all	Includes those proteins that have been extracted by at least one of the filter criteria.

4 Click **Apply filter** to view the results of the filtering in the heat map below.

If the result is not satisfactory, it is possible to add more/or remove protein and spot map filter criteria and click **Apply Filter** again. This procedure can be repeated until you are satisfied with the filters.

5 Click **Create set**. The **Create set** dialog is displayed.



Number of proteins: 170
Number of spot maps: 8

6 Enter a name for the set. If required, enter a comment on the set and select a color representing the set by clicking the color button and choosing the appropriate color.

Tip! *Different colors for the sets facilitates the analysis of the results in the results step.*

7 Click **Create** to create the set. The set will be added to the workspace but the previous set will still be displayed in the **Calculations**/**Results/Interpretation** window.

It is possible to create more sets, or to go to the **Calculations**/**Results/ Interpretation** step.

12.1.3 Protein filter criteria

When filtering a set it is possible to use:

- **General filter criteria**
  These criteria are used to remove missing values, filter the data using standard deviation and/or filter the data using the expression value range (max-min).
  These criteria are also available in the Manual Base Set Creation dialog used when creating the base set manually.

- **Calculation filter criteria**
  These criteria are used to filter the set based on the calculation results. The differential expression analysis calculation methods can be used as a protein filter criteria. If for example a Student's T-test was performed, it will appear in the drop-down list for filter criteria and a p-value for filtering can be entered. All values are numerical.

The different filter criteria are listed in Table 12-1 and Table 12-2 .

Criteria	Value	Description
% of spot maps where protein is present	Numerical (%)	***Tip!*** *Use this criteria to remove proteins that have many missing values among the spot maps.* Choose this criteria to include only those proteins that exist in a certain amount of spot maps in the data set. For example, if **>=** 80% is entered in the **Value** field, only proteins that have an expression value in >=80% of the spot maps (missing values are < 20%), will be included by the filter.

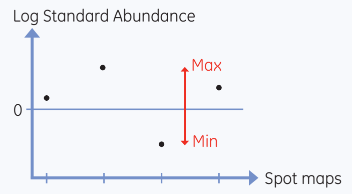
Criteria	Value	Description
% of exp. groups where protein is present	Numerical (%)	***Tip!*** *Use this criteria to remove proteins that have many missing values among the experimental groups.* Choose this criteria to include only those proteins that exist in a certain amount of experimental groups in the data set. For example, if >= 80% is entered in the **Value** field, only proteins that exist in >=80% of the experimental groups, will be included by the filter.
Standard deviation of log std. abundance	Numerical (range: 0-5)	Choose this criteria to include only those proteins with certain standard deviations. The standard deviation is a measure of the data spread and has the same unit as the observations (log standard abundance).
Log std. abundance difference	Numerical (>0)	Choose this criteria to only include proteins with a certain log standard abundance difference (**Max-Min** difference), i.e. proteins that have large expression differences among the spot maps. 

Table 12-1. General filter criteria.

Criteria	Description
Average Ratio **Paired Average Ratio**	Use these criteria to filter the set based on the results of the Average Ratio calculation. For example, if > 2 is entered in the **Value** field, only proteins with a 2-fold change (up-regulation) will be included by the filter. If entering < -2 only proteins with a 2-fold change (down-regulation) will be included by the filter. If a paired test was performed, the Paired Average Ratio will appear.

Criteria	Description
Student's T-test **Paired Student's T-test**	Use these criteria to filter the set based on the results of the Student's T-test calculation. For example, if **<** 0.01 is entered in the **Value** field, only proteins that received a p-value < 0.01 in the Student's T-test calculation will be included by the filter in the new set. If a paired test was performed, the **Paired Student's T-test** will appear. ***Note:*** *The calculation must have already been performed to appear in the Filter criteria list*.
One-Way ANOVA **RM One-Way ANOVA**	Use these criteria to filter the set based on the results of the **One-Way ANOVA** calculation. For example, if **<** 0.01 is entered in the **Value** field, only proteins that received a p-value < 0.01 in the **One-Way ANOVA** calculation will be included by the filter in the new set. If a paired test was performed, the **RM One-Way ANOVA** will appear. ***Note:*** *The calculation must have already been performed to appear in the Filter criteria list*.
Two-Way ANOVA, condition 1 **Two-Way ANOVA, condition 2** **Two-Way ANOVA, condition interaction**	Use these criteria to filter the set based on the results of the **Two-Way ANOVA** calculation. The **Two-Way ANOVA** calculation gives three p-values: **Two-Way ANOVA Condition 1**, **Two-Way ANOVA Condition 2** and **Two-Way ANOVA Interaction**. All these criteria will be available and can be used for filtering. For example, if the criteria **Two-Way ANOVA Condition 1** is selected and **<** 0.01 is entered in the **Value** field, only proteins that received a p-value < 0.01 in the **Two-Way ANOVA Condition 1** calculation will be included by the filter in the new set. ***Note:*** *The calculation must have already been performed to appear in the Filter criteria list*.
RM Two-Way ANOVA, condition 1 **RM Two-Way ANOVA, condition 2** **RM Two-Way ANOVA, condition interaction**	These criteria will appear if a paired test was performed. See above for description.

Table 12-2. Calculation filter criteria.

12.1.4 Spot Map filter criteria

Criteria	Value	Description
% of proteins present in spot map	Numerical (%)	***Tip!*** *Use this criteria to remove spot maps that have many missing protein expression values.* Choose this criteria to only include spot maps containing a certain amount of spots. For example, if >= 80% is entered in the **Value** field, only spot maps with at least 80% protein values (<20% missing values) are included by the filter.
Remove unassigned spot maps	n/a	Choose this criteria to remove all unassigned spot maps.

Table 12-3. Spot map filter criteria.

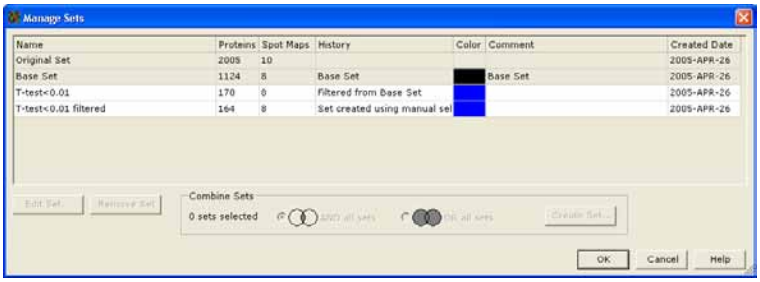
12.2 Managing sets

All available sets in the EDA workspace are listed in the **Manage set** dialog. It is possible to edit and remove sets and to create new sets by combining the available sets using the logical conditions **AND** and **OR**.

To manage sets:

1 Select **Tools:Manage sets...** in the menu bar.

The **Manage Sets** dialog is displayed.



Manage Sets

Name	Proteins	Spot Maps	History	Color	Comment	Created Date
Original Set	2005	10				2005-APR-26
Base Set	1124	8	Base Set	■	Base Set	2005-APR-26
T-test<0.01	170	8	Filtered from Base Set	🟦		2005-APR-26
T-test<0.01 filtered	164	8	Set created using manual sel	🟦		2005-APR-26

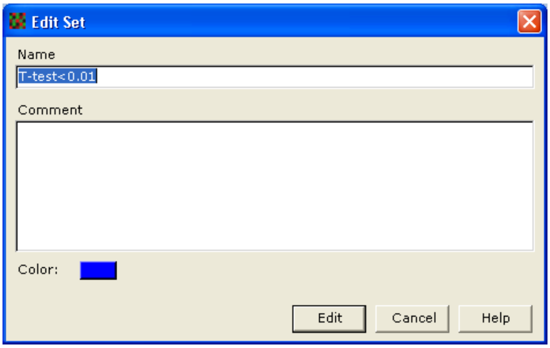
Combine Sets

0 sets selected

The original data set and base set are both displayed in gray and cannot be used to create new sets, edited or deleted. If you want to edit the base set you must re-create it in the **Setup** window. All created sets and calculations are lost when re-creating the base set.

2 *To edit a set:*

a. Select the set to be edited in the list and click **Edit set...**.The **Edit Set** dialog is displayed.



Edit Set

Name

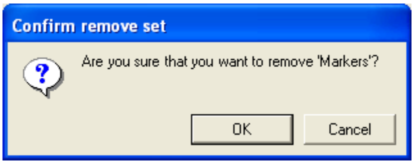
T-test<0.01

Comment

Color: 🟦

b. Edit the set as required (name, comment and/or color) and click **Edit**.

3 *To remove a set:* select the set(s) to remove from the list and click **Remove set**. A dialog appears asking you to confirm the removal of the set. Click **OK** to remove the set.


Confirm remove set

Are you sure that you want to remove 'Markers'?

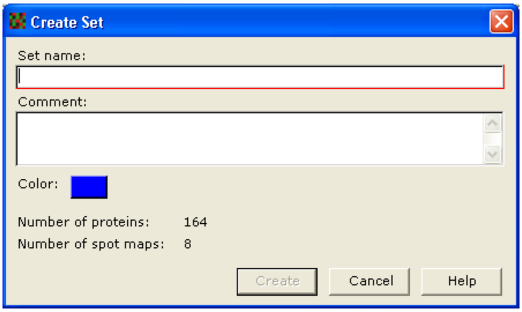
4 *To create a new set by combining sets from the list:*

a. Select the sets to be combined.

b. Select **AND all sets** to include only those proteins and spot maps that exist in both sets

Select **OR all sets** to include all proteins and spot maps that exists in at least one of the sets to be combined.

c. Click **Create Set...** . The **Create Set** dialog is displayed.


Create Set

Set name:

Comment:

Color: 🟦

Number of proteins: 164
Number of spot maps: 8

d. Enter a name for the set to be created. If required, enter a comment and change the color of the set.

The Proteins and Spot maps areas show how many proteins and spot maps will be included in the set.

e. Click **Create** to create the set.

5 In the **Manage Sets** dialog, click **OK** to apply the changes and close the dialog.

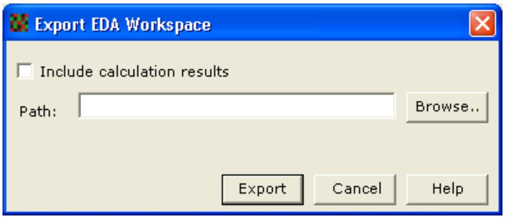
13 Exporting data from EDA

13.1 Overview

It is possible to export the EDA workspace to xml format and to copy data to the clipboard and then paste the data into Word, Excel or other applications.

13.2 Exporting the workspace to xml

1 Select **File:Export workspace...** in the menu bar. The **Export EDA workspace** dialog opens.


Export EDA Workspace

□ Include calculation results

Path:

2 Check the **Include calculation results** box to include calculations, settings and results (not graphs). If this box is not checked, only setup information (i.e. workspace name, spot maps, proteins, log standard abundance values, experimental groups and sets) are exported.

3 Enter the file path of the file to be exported in the **Path** field. The file path shows where the xml file will be saved and the file name after export.

Alternatively, click **Browse** and select the location of the file to export.

4 Click **Export** to export the EDA workspace to an xml file.

13.3 Copy data in EDA

Graphs and tables can be copied in EDA by clicking the appropriate graph/table and selecting **Edit:Copy** in the menu bar or using the shortcut **Ctrl + C**.

The copied data can then be pasted into, for example, reports in Word or Excel.

14 Tutorials Introduction

14.1 Scope of tutorials

The following tutorials are aimed at introducing the functionality of DeCyder EDA software within the context of an actual experiment. The tutorials have been designed to be step-by-step guides utilizing tutorial files. The two tutorial chapters cover different aspects of the software suite. They are both self-contained and can be undertaken independently.

To assist the user, each tutorial includes a completed version of the EDA file, which the tutorial is designed to generate. These files all include the word **finished** in their names.

The tutorials described below introduce the concepts and functionality of DeCyder EDA module. It is therefore recommended that these tutorials are performed first to gain a preliminary understanding of the software.

14.1.1 Tutorial I - Identify spots for picking and import MS data

This tutorial demonstrates how to perform pattern analyses of differentially expressed proteins and how to select proteins of interest from which to generate a pick list and import MS Data. The protein patterns of brain tissues from two mouse strains, wildtype (wt) and mice with the gene NDST-1 knocked out (mutant), were analyzed using EDA.

See Chapter 15, Tutorial I - Identify spots for picking and import MS data.

14.1.2 Tutorial II - Classification of ovarian cancer biopsies

This tutorial demonstrates how a dataset with already classified biopsy material (normal, benign and malignant) from a human ovarian cancer study can be used to find biomarkers that discriminate between the different classes. This tutorial also demonstrates how to create a classifier that can classify biopsy material of unknown class.

See chapter 16, Tutorial II - Classification of ovarian cancer biopsies.

14.2 Tutorial files

The tutorial files are provided on a DVD. The tutorial projects should be imported into the database when installing the software, see *DeCyder 2D Differential Analysis Software Installation Guide* for instructions about how to perform the import.

Tutorial I contains the files listed in Table 14-1.

File/Folder	Type of file	Description
EDA Tutorial I	BVA workspace	The EDA workspace will be created from this BVA workspace.
EDA Tutorial I finished	EDA workspace	This file shows the results of Tutorial I.
EDA Tutorial I pick list finished	Text file (*.txt)	This file contains the pick list, which is also created in the tutorial.
EDA Tutorial I MS data	Folder with search results from Mascot (*.dat files)	This folder contains the *.dat files with the search results from Mascot.

Table 14-1. EDA Tutorial I files/folder.

Tutorial II contains the files listed in Table 14-2.

File	Type of file	Description
EDA Tutorial II start	EDA workspace	A copy of this workspace must be created before starting to work with this tutorial. See section 16.4, Copy the tutorial file to your own project for more information.
EDA Tutorial II finished	EDA workspace	This file shows the results of Tutorial II.

Table 14-2. EDA Tutorial II files.

15 Tutorial I - Identify spots for picking and import MS data

15.1 Objective

This tutorial describes how to perform PCA and pattern analyses of differentially expressed proteins and how to generate a pick list and import MS data. Data from a single BVA workspace is used to illustrate how expression data can be evaluated and visualized further in EDA compared to BVA.

Also covered is the central concept of how to create sets.

The purpose of this tutorial is to teach how to:

- Perform differential expression analysis (T-test) in EDA
- Create sets
- Perform PCA
- Perform pattern analysis (hierarchical clustering)
- Generate a pick list in BVA from a set in EDA
- Import MS data

15.2 Experiment overview

15.2.1 Introduction

The protein patterns of brain tissues from two mouse strains, wildtype (wt) and mice with the gene NDST-1 knocked out (mutant), will be analyzed using EDA.

The purpose of the experiment is to find which proteins differ in expression between the wildtype and mutant mice. In EDA this can be performed in a more extended way compared to BVA because:

- In addition to the differential expression analysis, a better overview of the data can be obtained by performing PCA and hierarchical clustering, revealing patterns of sub-groups among the proteins and spot maps.
- Also, biological interpretation of the proteins of interest that are found can be performed.
- Sets and sub-sets of proteins and/or spot maps can easily be created for further calculations and analyses.

- Filtering of the data, using different criteria, facilitates the extraction of proteins of interest from different points of view

15.2.2 Experimental design

Table 15-1 gives an overview of the experimental design in EDA with experimental groups, colors for the experimental groups, mouse strains and numbers of replicates. The unassigned group contains the standard spot map set to Master, which is removed when creating the base set on which to perform calculations.

Experimental group	Color	Mouse strain	# mice	#replicates/ mouse
1-wt	Orange	wt	3	2
2-mutant	Blue	mutant	3	2

Table 15-1. Experimental groups in EDA.

15.2.3 Basic work already performed

Pre-processing of the gels in DIA and the BVA module have been performed, giving one BVA workspace (brain) with two experimental groups (wt and mutant). When starting to work with this tutorial, this BVA workspace is used to create the EDA workspace.

15.3 EDA workflow overview

1 Start EDA.

2 Setup and save the EDA workspace. Setup includes importing BVA workspaces and creating the base set.

3 Perform differential expression analysis

4 Filter the results and create a new set

5 Perform PCA

6 Perform pattern analysis (hierarchical clustering)

7 Create a set with proteins to pick

8 Generate pick list

9 Import MS data

Sections 15.4-15.11 describe the different steps to perform in detail. Fig. 15-1 outlines an overview of the experiment.

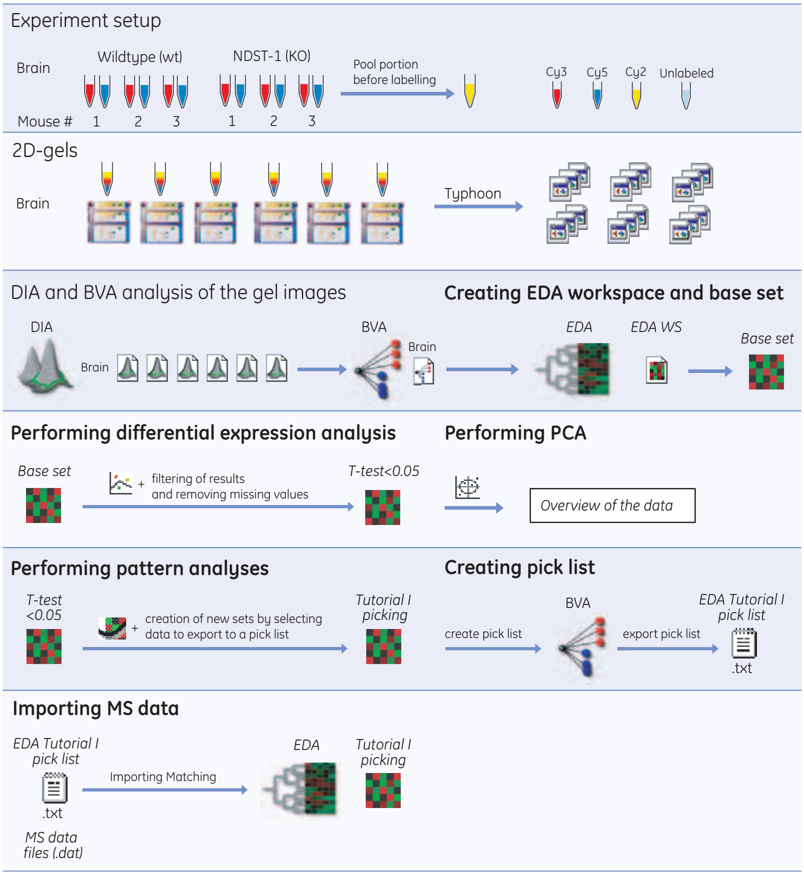
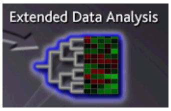


Fig 15-1. Overview of the experiment. Experiment setup, DIA and BVA analysis of the Typhoon images have already been performed and are not included in the tutorial.

15.4 Start EDA

1 Start DeCyder 2D Software, see section 2.3.

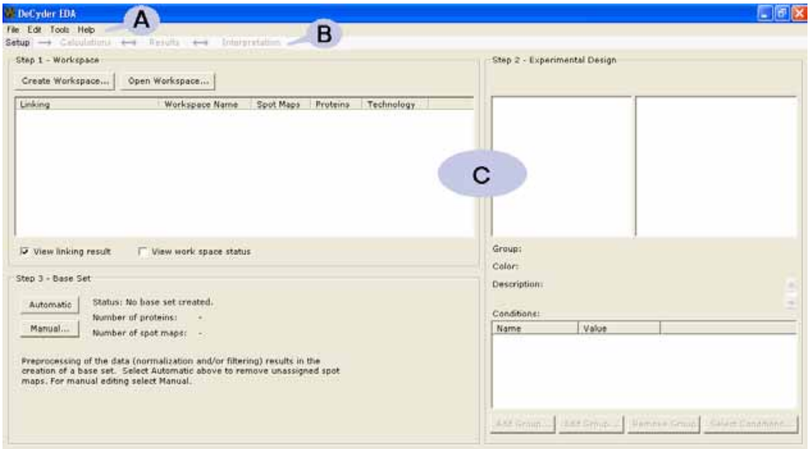


2 Click the Extended Data Analysis (EDA) icon in the DeCyder 2D main window.

3 EDA will open displaying the DeCyder EDA main screen, which is divided into three areas:

- menu bar (A)
- workflow area (B)
- work area (C)

Depending on the currently selected step in the workflow area, the work area will appear different. In the beginning, the first step in the workflow, **Setup**, is selected and the **Setup** window is displayed in the work area.



Status: No base set created.
Number of proteins: -
Number of spot maps: -

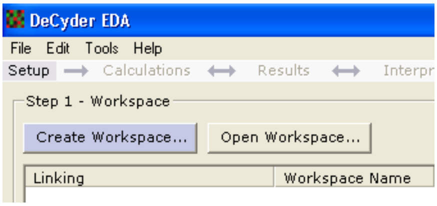
Preprocessing of the data (normalization and/or filtering) results in the creation of a base set. Select Automatic above to remove unassigned spot maps. For manual editing select Manual.

15.5 Set up and save the EDA workspace

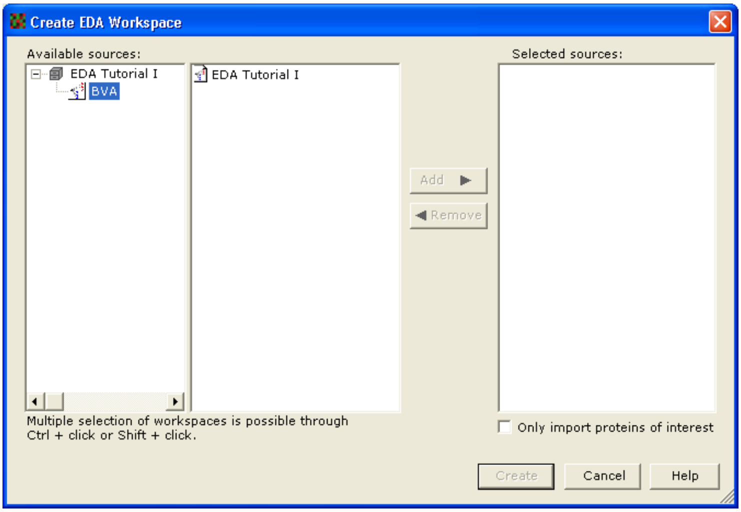
The EDA workspace is set up by importing the BVA workspaces and creating the base set.

15.5.1 Create the EDA workspace

1 Click **Create workspace...** in the **Step 1 - Workspace** area of the **Setup** window.



The **Create EDA Workspace** dialog opens.



Multiple selection of workspaces is possible through Ctrl + click or Shift + click.

2 Double-click the **EDA Tutorial I** project in the **Available Workspace(s)** area (in the left panel) and click on the BVA icon.

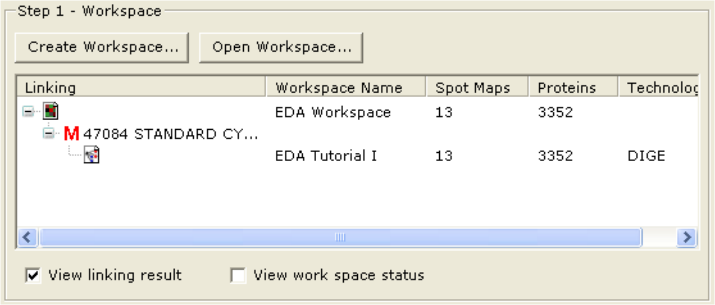
The BVA workspaces included in the project are shown to the right.

3 Select the **EDA Tutorial I** BVA workspaces and click **Add -->**.

The added BVA workspace is displayed in the **Selected sources** area (right panel).

4 Click **Create** to create the EDA workspace.

5 The created EDA workspace is displayed in the **Step 1 - Workspace** area of the **Setup** window.



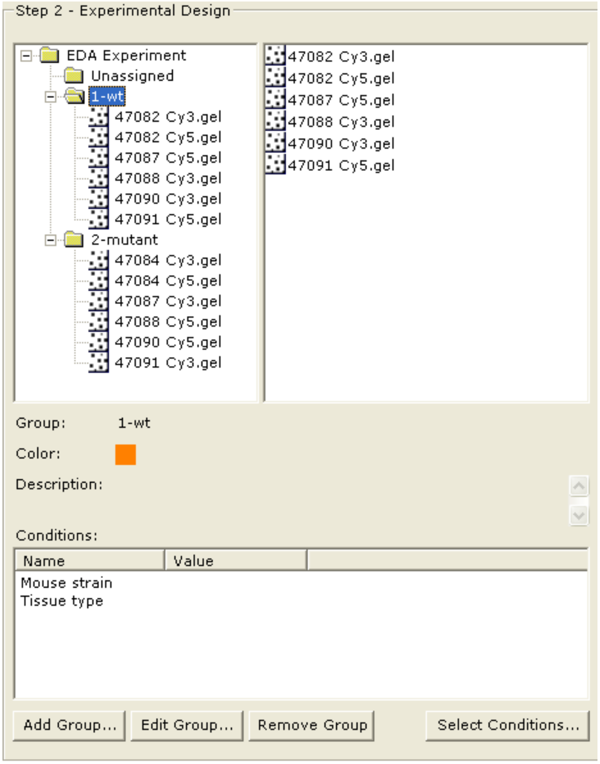
Step 1 - Workspace

Linking	Workspace Name	Spot Maps	Proteins	Technolog
	EDA Workspace	13	3352	
M 47084 STANDARD CY...				
	EDA Tutorial I	13	3352	DIGE

☑ View linking result □ View work space status

Basic information about the BVA workspace that the EDA workspace was created from (name, the number of spot maps, proteins and linking) is displayed.

6 The experimental design in the BVA workspace is transferred to EDA and displayed in **Step 2 - Experimental Design** area of the **Setup** window.



Step 2 - Experimental Design

```
EDA Experiment
    Unassigned
    1-wt
        47082 Cy3.gel
        47082 Cy5.gel
        47087 Cy5.gel
        47088 Cy3.gel
        47090 Cy3.gel
        47091 Cy5.gel
    2-mutant
        47084 Cy3.gel
        47084 Cy5.gel
        47087 Cy3.gel
        47088 Cy5.gel
        47090 Cy5.gel
        47091 Cy3.gel
```

```
47082 Cy3.gel
47082 Cy5.gel
47087 Cy5.gel
47088 Cy3.gel
47090 Cy3.gel
47091 Cy5.gel
```

Group: 1-wt

Color: 🟧

Description:

Conditions:

Name	Value
Mouse strain	
Tissue type	

The experimental design is correct and does not need to be edited.

15.5.2 Create the base set

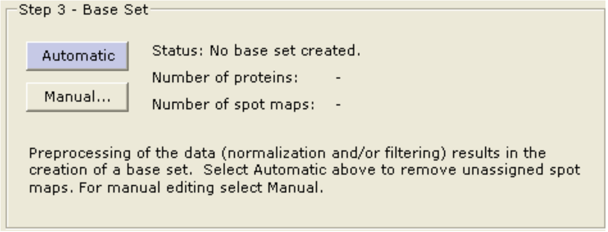
A base set must always be created before any analyses can be performed. When the base set has been created, the rest of the steps in the workflow area become activated and new sets can be created and calculations can be performed.

The base set can be created either manually or automatically. In this tutorial the base set is created automatically. Spot maps located in the **Unassigned** folder (Masters and spot maps that are not assigned to an experimental group) are then removed from the dataset.

Note: *Missing values will be excluded later, before performing PCA.*

To create the base set:

1 Click **Automatic** in the **Step 3 - Base Set** area.

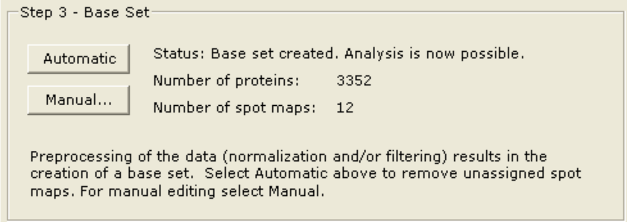


Status: No base set created.
Number of proteins: -
Number of spot maps: -

Preprocessing of the data (normalization and/or filtering) results in the creation of a base set. Select Automatic above to remove unassigned spot maps. For manual editing select Manual.

2 The base set is created. During the creation, a progress bar is displayed.

3 When the base set has been created, the **Base set created**- **Calculation is now possible** status is displayed in the **Status** field. The number of proteins and spot maps included in the base set is also displayed.



Status: Base set created. Analysis is now possible.
Number of proteins: 3352
Number of spot maps: 12

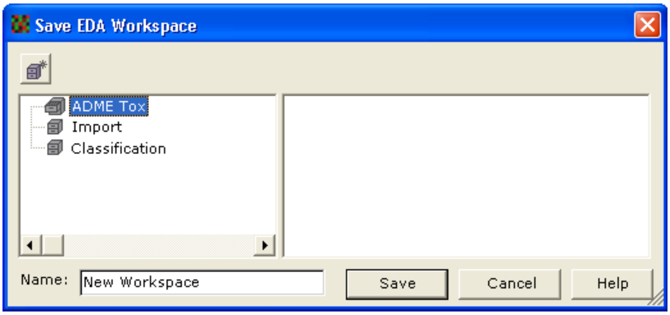
Preprocessing of the data (normalization and/or filtering) results in the creation of a base set. Select Automatic above to remove unassigned spot maps. For manual editing select Manual.

15.5.3 Save the workspace

When the base set has been created it is recommended to save the EDA workspace.

To save the workspace:

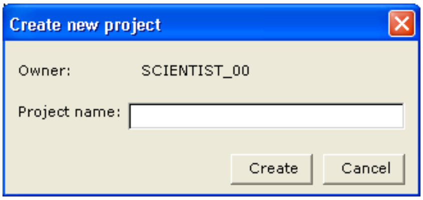
1 Select **File:Save Workspace** in the menu bar. The **Save EDA workspace** dialog is displayed.



Name: New Workspace

2 Click the **New project** icon to create a new project in the database in which to save your personal work on tutorial files.

3 The Create new project dialog is displayed.



Owner: SCIENTIST_00

Project name:

4 Enter a name for the project and click **OK** to create the project and return to the **Save EDA workspace** dialog. The created project will be selected in the **Save EDA workspace** dialog.

5 Enter your name as the name for the workspace in the **Name** field.

6 Click **Save**.

15.6 Perform differential expression analysis

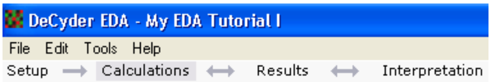
The EDA workspace has been set up and saved and the base set created.

Differential expression analysis is now going to be performed on the base set to find significantly differentially expressed proteins when comparing brain tissues from wildtype and mutant mice.

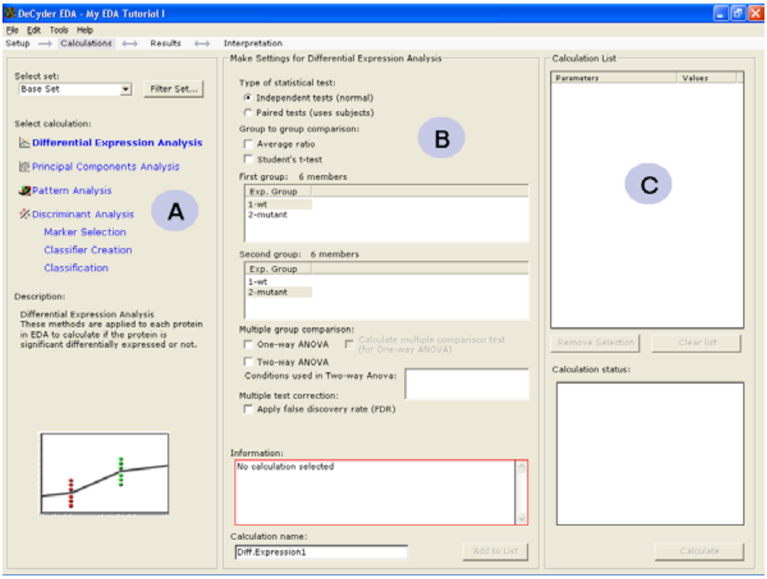
The aim is to find proteins that differ in expression when comparing the two groups.

15.6.1 Set up the differential expression analysis calculations

1 Click **Calculations** in the workflow area.



The **Calculations window** opens displaying the settings for differential expression analysis by default.



Select set: Base Set

Select calculation: Differential Expression Analysis

Description:
Differential Expression Analysis
These methods are applied to each protein in EDA to calculate if the protein is significant differentially expressed or not.

Type of statistical test:
● Independent tests (normal)
○ Paired tests (uses subjects)

Group to group comparison:
□ Average ratio
□ Student's t-test

First group: 6 members

Exp. Group
1-wt
2-mutant

Second group: 6 members

Exp. Group
1-wt
2-mutant

Multiple group comparison:
□ One-way ANOVA □ Calculate multiple comparison test (for One-way ANOVA)
□ Two-way ANOVA
Conditions used in Two-way Anova:

Multiple test correction:
□ Apply false discovery rate (FDR)

Information:
No calculation selected

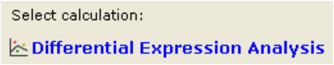
Calculation name: Diff.Expression1

Calculation status:

Fig 15-2. Calculations window. The window is divided into three areas: **Calculations**, **A** (where the set on which to perform calculations and the type of calculation are selected), Make Settings, **B** (where settings for the selected calculation in **A** are entered) and **Calculation List**, **C** (where the added calculations are listed and can be calculated).

2 By default, the **Differential Expression Analysis** is selected in the **Select**

calculation area and the settings for the calculation are displayed to the right.


Select calculation:
Differential Expression Analysis

3 Enter the following settings:

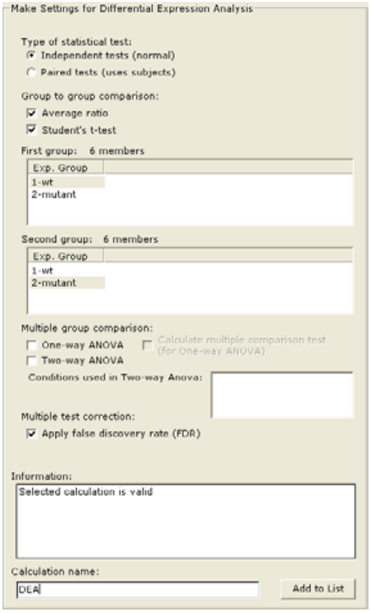
a. Choose **Independent tests (normal)** in the **Type of statistical tests** area.

b. Check the **Average ratio** and **Student's T-test** boxes in the **Group to group comparison** area.

c. Select **1-wt** in the **First group** area and **2-mutant** in the **Second group** area.

d. Check the **Apply false discovery rate (FDR)** box in the **Multiple test correction** area.

e. Leave all other boxes unchecked and enter *"DEA"* in the **Calculation name** field.


Make Settings for Differential Expression Analysis

Type of statistical test:
● Independent tests (normal)
○ Paired tests (uses subjects)

Group to group comparison:
☑ Average ratio
☑ Student's t-test

First group: 6 members

Exp. Group
1-wt
2-mutant

Second group: 6 members

Exp. Group
1-wt
2-mutant

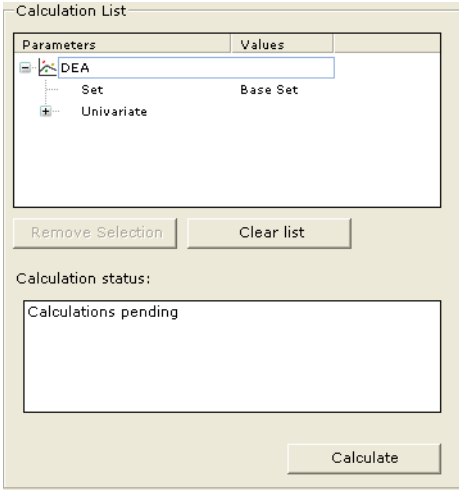
Multiple group comparison:
☐ One-way ANOVA
☐ Two-way ANOVA
☐ Calculate multiple comparison test (for One-way ANOVA)
Conditions used in Two-way Anova:

Multiple test correction:
☑ Apply false discovery rate (FDR)

Information:
Selected calculation is valid

Calculation name:
DEA

4 Click **Add to List** to add the calculation to the **Calculation List** to the right.


Calculation List

Parameters	Values
DEA	
Set	Base Set
Univariate	

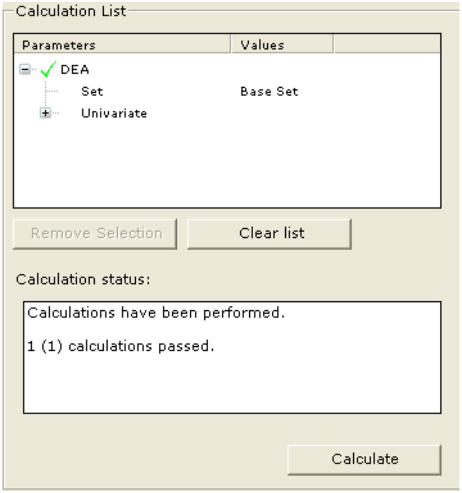
Calculation status:
Calculations pending

5 Click **Calculate** to start the calculation.

During calculation the status of each calculation is indicated by an icon in front of the calculation and the progress of the calculations are displayed in a progress bar.

Icon	Description
🟢	The calculation is in progress.
✓	The calculation has successfully finished.
✗	The calculation has been cancelled.
🔴	The calculation has failed.

6 When the calculation has finished the status of the calculation is also displayed in the **Calculation Status** area.



Calculations have been performed.

1 (1) calculations passed.

15.6.2 Filter the results and create a new set

The differential expression analysis calculation on the **base set** has been performed.

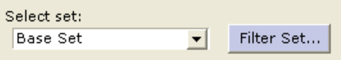
The following filtering will be performed:

- a filter criteria extracting significantly differentially expressed proteins (with p-value < 0.05) will be added
- a filter criteria removing protein values missing in more than 80% of the spot maps will be added (too many missing values will affect the PCA calculation)

When the filters have been applied, a new set containing the filtered data will be created.

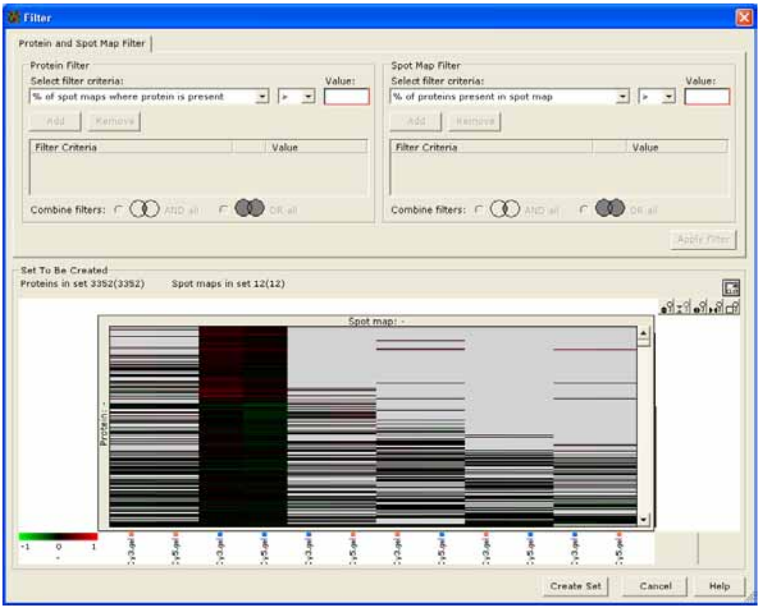
Filter the results and create the T-test<0.05 set

1 Still in the **Calculations** step, click **Filter Set...**.



Select set: Base Set

The **Filter dialog** is displayed.



Protein Filter
Select filter criteria: % of spot maps where protein is present
Value:

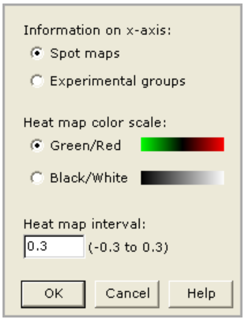
Spot Map Filter
Select filter criteria: % of proteins present in spot map
Value:

Filter Criteria	Value

Combine filters: ○ AND all ○ OR all

Set To Be Created
Proteins in set 3352(3352)   Spot maps in set 12(12)

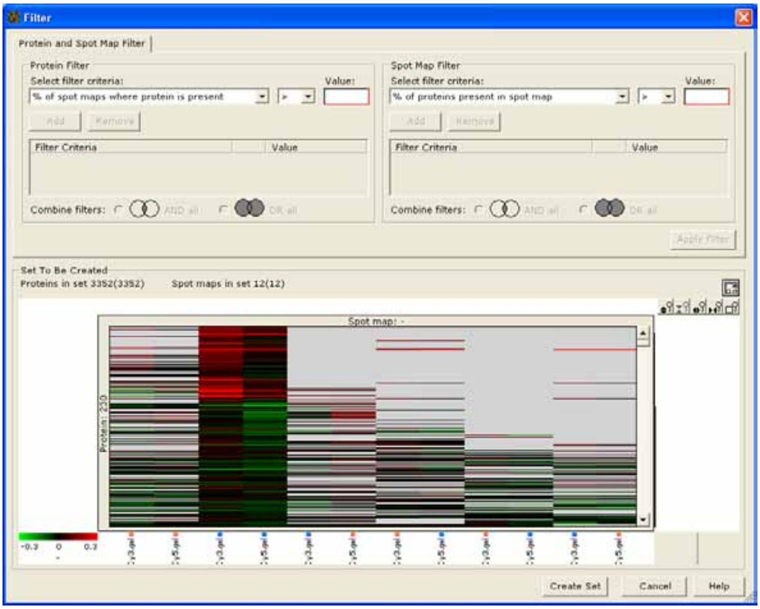
Click the **Settings** icon to display the **Heat map settings** pop-up dialog, change the heat map interval to 0.3 and click OK. The heat map is updated.



Information on x-axis:
● Spot maps
○ Experimental groups

Heat map color scale:
● Green/Red
○ Black/White

Heat map interval:
0.3 (-0.3 to 0.3)



Filter

Protein and Spot Map Filter

Protein Filter
Select filter criteria: % of spot maps where protein is present > Value:

Spot Map Filter
Select filter criteria: % of proteins present in spot map > Value:

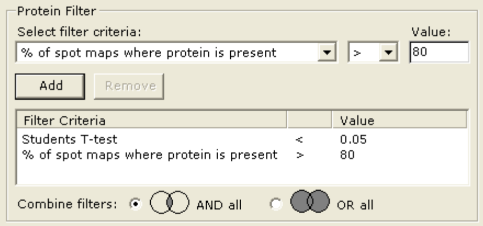
Set To Be Created
Proteins in set 3352(3352)   Spot maps in set 12(12)

2 To extract all proteins with a p-value < 0.05 (from the Student's T-test calculation), and remove proteins with more than 80% missing values, select the protein filter as follows:

a. Select the filter criteria **Student's T-test**.

b. Choose the operator **<** and enter the value **0.05**.

c. Click **Add**. The filter criteria is added to the list below.



Protein Filter

Select filter criteria: % of spot maps where protein is present > Value: 80

Filter Criteria		Value
Students T-test	<	0.05
% of spot maps where protein is present	>	80

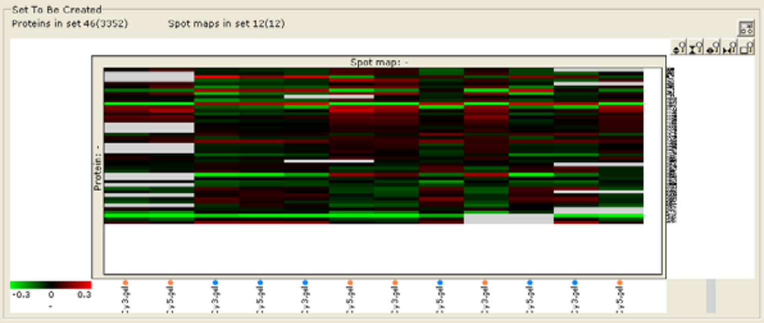
Combine filters: ● AND all ○ OR all

d. Select the filter criteria **% of spot maps where protein is present**.

e. Choose the operator **>**, enter the value **80** and click **Add**.

f. Make sure that the **AND all** radio button is selected in the **Combine filters** field.

3 Click **Apply Filter** to apply the filter criteria on the base set. The heat map will be updated and information on the number of remaining spot maps and proteins are displayed in the **Set To Be Created** area.

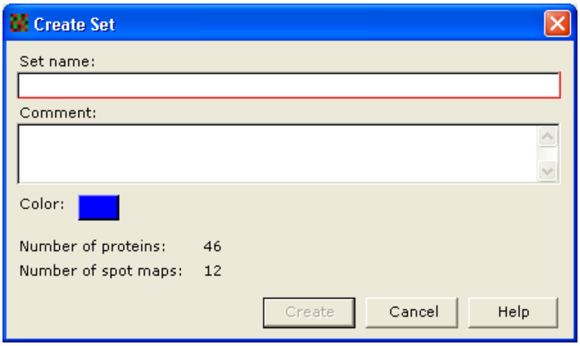


Set To Be Created

Proteins in set 46(3352)

Spot maps in set 12(12)

4 Click **Create set**. The **Create set** dialog is displayed.



Create Set

Set name:

Comment:

Color:

Number of proteins: 46

Number of spot maps: 12

5 Type in *"T-test<0.05"* in the **Set name** field and enter *"Missing values removed and proteins with p-value <0.05 extracted"* in the **Comment** field.

6 Click **Create**. The **T-test<0.05** set is created. It will be added to the **Select set** list.

15.7 Perform PCA

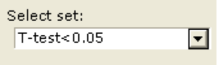
A set with significantly differentially expressed proteins and where missing values have been removed has been created (**T-test<0.05**). PCA on this set is now going to be performed. Two PCA calculations are going to be set up, one on proteins versus spot maps and one on spot maps versus proteins.

In the PCA calculations, it is possible to see:

- Which proteins lie outside of a 95% significance level in their expression
- The relation between proteins and spot maps
- If replica spot maps are grouped together

PCA is mainly performed to obtain an overview of the data and to check that the data looks OK (e.g. there are no spot map and/or protein outliers).

15.7.1 Set up and calculate the PCA calculations



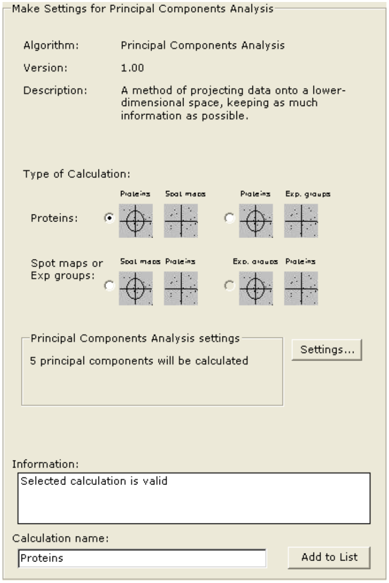
Select set:
T-test<0.05

1 In the **Calculations** step, select the **T-test<0.05** set in the **Selected set** pop-up dialog.

2 Select **Principal Component Analysis** in the **Select calculation** area. The settings for the calculation are displayed to the right.

Principal Components Analysis

3 Enter the following settings for the PCA calculation on proteins-spot maps:



Make Settings for Principal Components Analysis

Algorithm: Principal Components Analysis

Version: 1.00

Description: A method of projecting data onto a lower-dimensional space, keeping as much information as possible.

Type of Calculation:

Proteins: ● Proteins / Spot maps ○ Proteins / Exp. groups

Spot maps or Exp groups: ○ Spot maps / Proteins ○ Exp. groups / Proteins

Principal Components Analysis settings

5 principal components will be calculated

Information:

Selected calculation is valid

Calculation name: Proteins

a. In the **Type of Calculation** area, choose the left radio button in the **Proteins** area.

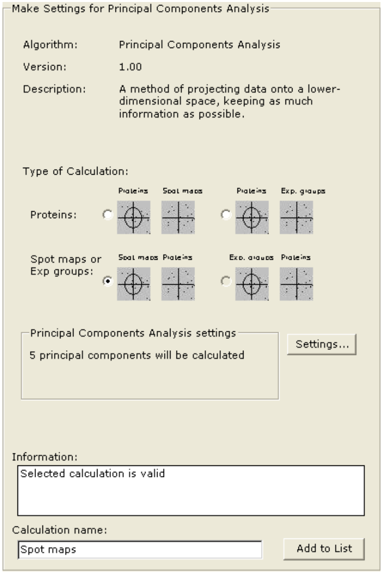
This setting will give an overview of the data, with proteins in the score plot (left plot) and spot maps in the loading plot (right plot).

b. Use the default settings displayed in the **Principal Component Analysis settings** area.

c. Type in *"Proteins "* in the **Calculation name** field.

4 Click **Add to List** to add the calculation to the **Calculation List** to the right.

5 Enter the following settings for the PCA calculation on spot maps-proteins:


Make Settings for Principal Components Analysis

Algorithm: Principal Components Analysis

Version: 1.00

Description: A method of projecting data onto a lower-dimensional space, keeping as much information as possible.

Type of Calculation:

Proteins: ○ Proteins / Spot maps ○ Proteins / Exp. groups

Spot maps or Exp groups: ● Spot maps / Proteins ○ Exp. groups / Proteins

Principal Components Analysis settings

5 principal components will be calculated

Information:

Selected calculation is valid

Calculation name: Spot maps

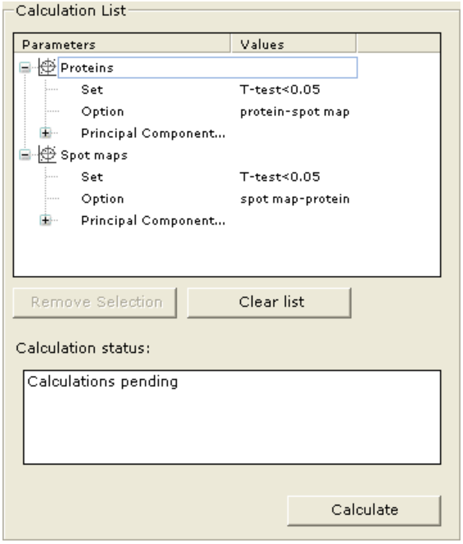
a. In the **Type of Calculation** area, choose the left radio button in the **Spot maps or Exp groups** area.

   This setting will give an overview of the data, with spot maps in the score plot (left plot) and proteins in the loading plot (right plot).

b. Use the default settings displayed in the **Principal Component Analysis settings** area.

c. Type in *"Spot maps"* in the **Calculation name** field.

6 Click **Add to List** to add the calculation to the **Calculation List** to the right.


Calculation List

Parameters	Values
Proteins	
Set	T-test<0.05
Option	protein-spot map
Principal Component...	
Spot maps	
Set	T-test<0.05
Option	spot map-protein
Principal Component...	

Calculation status:

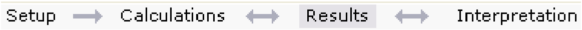
Calculations pending

7 Click **Calculate** to start the calculation.

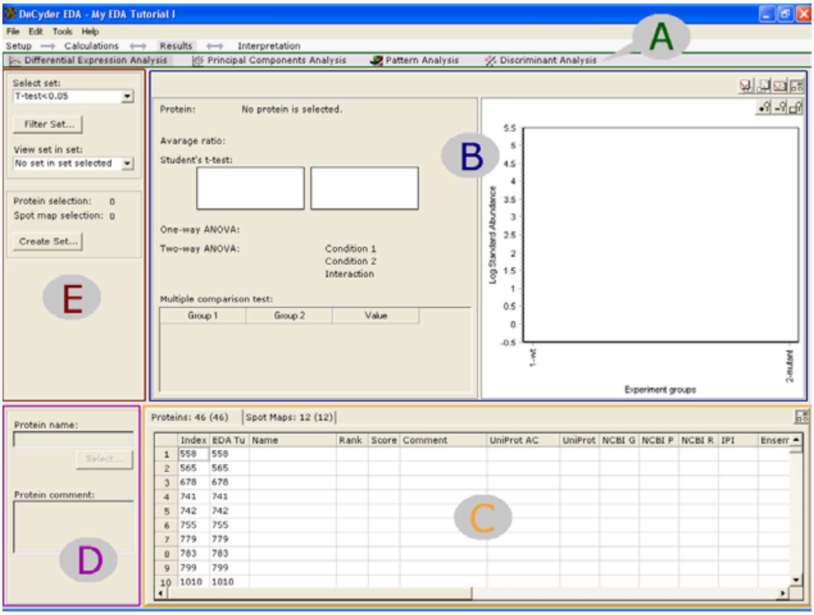
During calculation the status of the calculation is indicated by an icon in front of the calculation and the progress of the calculation is displayed by a progress bar.

15.7.2 View and analyze the results of the PCA calculation

1 Select the **Results** step in the workflow area.



The **Results** window is displayed.



Select set: T-test<0.05

View set in set: No set in set selected

Protein selection: 0

Spot map selection: 0

Protein: No protein is selected.

Avarage ratio:

Student's t-test:

One-way ANOVA:

Two-way ANOVA: Condition 1 Condition 2 Interaction

Multiple comparison test:

Group 1	Group 2	Value

Protein name:

Protein comment:

Proteins: 46 (46) | Spot Maps: 12 (12)

	Index	EDA Tu	Name	Rank	Score	Comment	UniProt AC	UniProt	NCBI G	NCBI P	NCBI R	IPI	Enserr
1	558	558											
2	565	565											
3	678	678											
4	741	741											
5	742	742											
6	755	755											
7	779	779											
8	783	783											
9	799	799											
10	1010	1010											

The results window is divided into five main areas.

- **Results bar (A)**
  In this area, select the analysis results to display in the results view (B) and protein/spot map table (C) by clicking on the appropriate analysis.

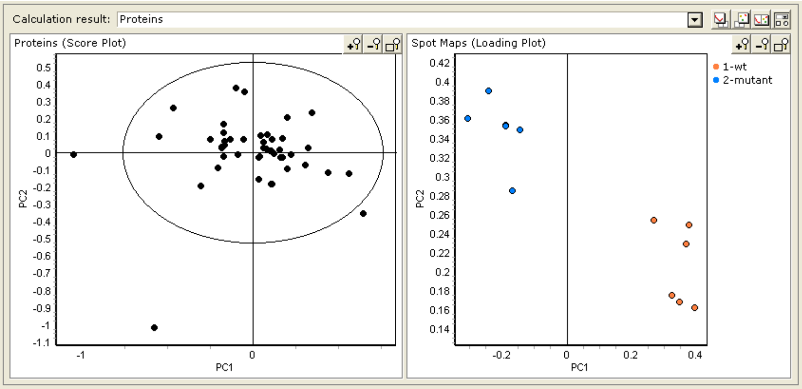
- **Results view (B)**
  Shows details of results for the selected protein/spot map in the protein/spot map table for the current calculation.

- **Protein/Spot map table (C)** and **Protein/spot map details area (D)**
  The Protein and Spot map tables (C) show information for all proteins and spot maps in a table format. When highlighting a protein/spot map in the tables or in the results view, details for the selected protein/spot map will be displayed in the protein/spot map details areas (D).

- **Set area (E)**
  In this area new data sets can be created by selecting data directly from the results view and/or protein/spot map table or by filtering the data.

2 Click **Principal Component Analysis** in the Results bar.

3 The result of the *Proteins* PCA calculation is displayed in the Results view. The **Calculation result** field shows the name of the calculation for which results are displayed.



Calculation result: Proteins

Analyze the results as follows:

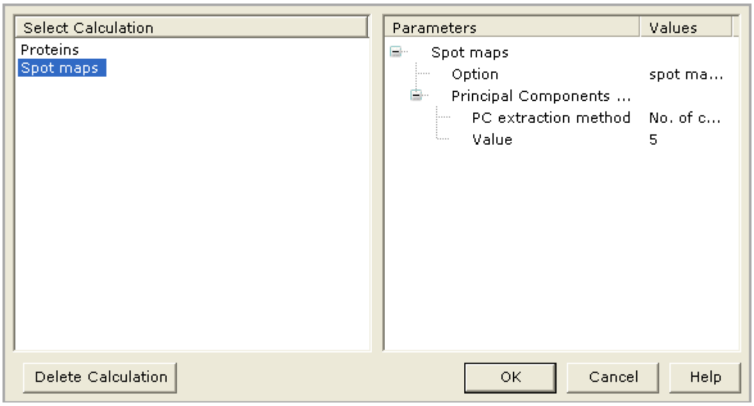
- **Score plot**
  The score plot (left plot) shows an overview of the proteins. The ellipse represents a 95% significance level.

  Three proteins lie outside of the ellipse and are outliers. Protein outliers can either be very strongly differentially expressed proteins or mismatched spots. In this case, the outliers have been checked in BVA and they are strongly differentially expressed proteins.

- **Compare the two plots**
  It is possible to perform a **rough** comparison of the relationship between proteins and spot maps and to **estimate** which proteins are up- or down-regulated in the different spot maps. Proteins and spot maps located in the corresponding quadrants have a correlation.

  For example, proteins in the upper-left quadrant of the score plot are probably up-regulated in the blue spot maps but down-regulated in the orange spot maps. Proteins in the upper-right quadrant are probably up-regulated in the orange spot maps but down-regulated in the blue spot maps.

4 In the **Calculation result** field, select the results to display from the spot maps versus proteins PCA calculation by clicking the arrow and selecting the *Spot maps* calculation in the **Select calculation** pop-up dialog.



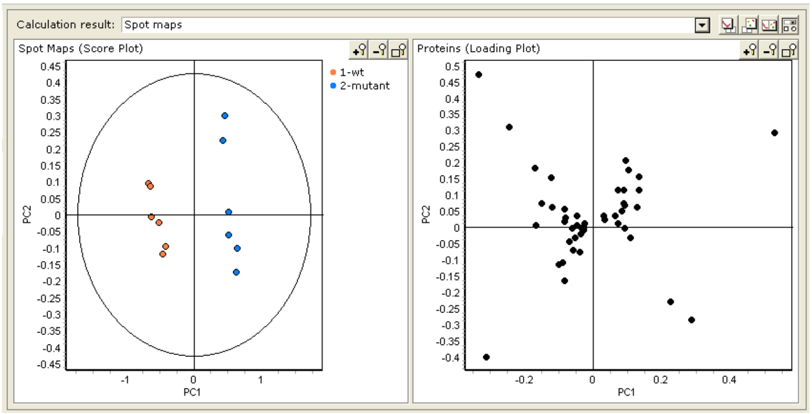
Select Calculation

Proteins

Spot maps

Parameters	Values
Spot maps	
Option	spot ma...
Principal Components ...	
PC extraction method	No. of c...
Value	5

5 The results for the calculation are displayed in the Results view.



Calculation result: Spot maps

Analyze the results as follows:

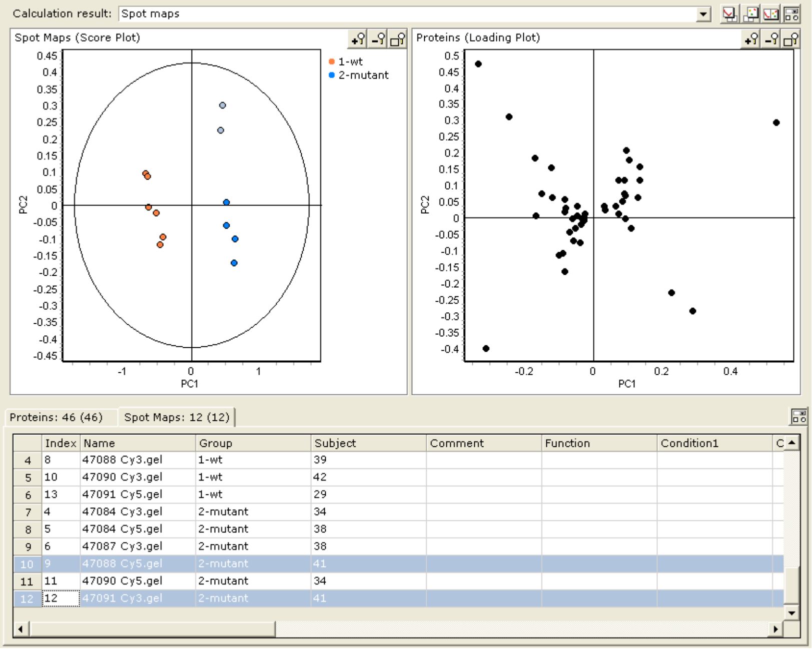
Score plot

The score plot (left plot) shows an overview of the spot maps. The ellipse represents a 95% significance level.

In the plot, the orange spot maps (wt) have been grouped together. In the case of the mutant mice (blue spot maps), two spot maps deviate from the other four but these two spot maps are grouped together.

If selecting the two spot maps in the score plot (press Ctrl + click) and selecting the spot map table at the bottom of the screen, the spot maps represent the replica spot maps for one mouse (subject 41).

This result indicates that we have a sub group within the mutant group which is an interesting observation but is not investigated further in this tutorial.



Calculation result: Spot maps

Proteins: 46 (46) | Spot Maps: 12 (12)

	Index	Name	Group	Subject	Comment	Function	Condition1	C
4	8	47088 Cy3.gel	1-wt	39				
5	10	47090 Cy3.gel	1-wt	42				
6	13	47091 Cy5.gel	1-wt	29				
7	4	47084 Cy3.gel	2-mutant	34				
8	5	47084 Cy5.gel	2-mutant	38				
9	6	47087 Cy3.gel	2-mutant	38				
10	9	47088 Cy5.gel	2-mutant	41				
11	11	47090 Cy5.gel	2-mutant	34				
12	12	47091 Cy3.gel	2-mutant	41				

6 An overview of the data has been produced. Proceed with section 15.8.

15.8 Perform pattern analysis

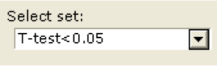
An overview of the data has been produced giving an initial view of the groupings and data relations.

The pattern analysis **Hierarchical clustering** will now be performed on the **T-test<0.05** set. This analysis is performed to investigate the expression patterns among the proteins and the overall expression patterns from the spot maps. It is possible to see if the replica spot maps are clustered together and to view the general protein patterns in the data set.

Hierarchical clustering is one of the most frequently used clustering algorithms. It is a method that combines or splits the data pairwise and thereby generates a treelike structure called a dendrogram. The dendrogram and the heat map are displayed together in the result. This analysis rearranges the data set into a new, better ordered data set. See Appendix E, Statistics and algorithms - Pattern Analysis for more information on Hierarchical clustering.

15.8.1 Set up the hierarchical clustering calculation

1 Select the **Calculations** step in the workflow area.


Select set:
T-test<0.05

2 In the **Calculations** step, select the **T-test<0.05** set in the **Select set** pop-up dialog.

3 Select **Pattern Analysis** in the **Select calculation** area. The settings for the calculation are displayed to the right.

4 Enter the following settings for the Pattern Analysis calculation on proteins:

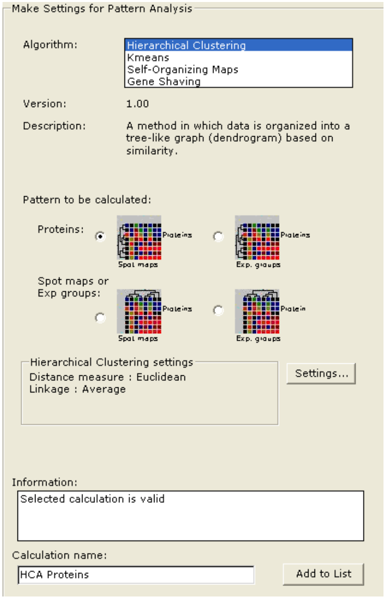
a. In the **Algorithm** area, select **Hierarchical clustering**

b. In the **Pattern to be calculated area**, choose the left radio button in the **Proteins** area.

c. Use the default settings displayed in the **Hierarchical Clustering settings** area.

d. Enter *HCA Proteins* in the **Calculation name** field.

5 Click **Add to List** to add the calculation to the **Calculation List** to the right.


Make Settings for Pattern Analysis

Algorithm: Hierarchical Clustering / Kmeans / Self-Organizing Maps / Gene Shaving

Version: 1.00

Description: A method in which data is organized into a tree-like graph (dendrogram) based on similarity.

Pattern to be calculated:

Proteins: ● ○

Spot maps or Exp groups: ○ ○

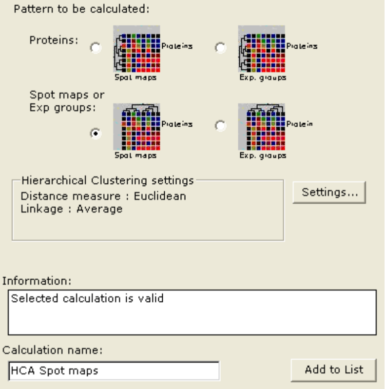
Hierarchical Clustering settings
Distance measure : Euclidean
Linkage : Average

Information:
Selected calculation is valid

Calculation name:
HCA Proteins

6 Repeat steps 4-5 to enter settings for the hierarchical clustering analysis on spot maps but:

- in step 4b, select the left radio button in the **Spot maps or Exp. groups** area
- in step 4d enter *HCA Spot maps* in the **Calculation name** field.


Pattern to be calculated:

Proteins: ○ ○

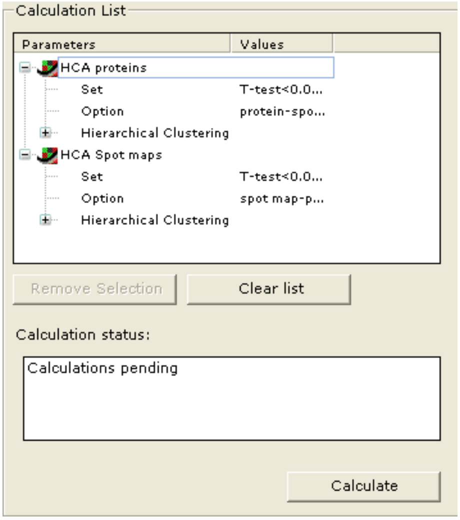
Spot maps or Exp groups: ● ○

Hierarchical Clustering settings
Distance measure : Euclidean
Linkage : Average

Information:
Selected calculation is valid

Calculation name:
HCA Spot maps

7 Click **Add to List** to add the calculation to the **Calculation List**.



Calculation List

Parameters	Values
HCA proteins	
Set	T-test<0.0...
Option	protein-spo...
Hierarchical Clustering	
HCA Spot maps	
Set	T-test<0.0...
Option	spot map-p...
Hierarchical Clustering	

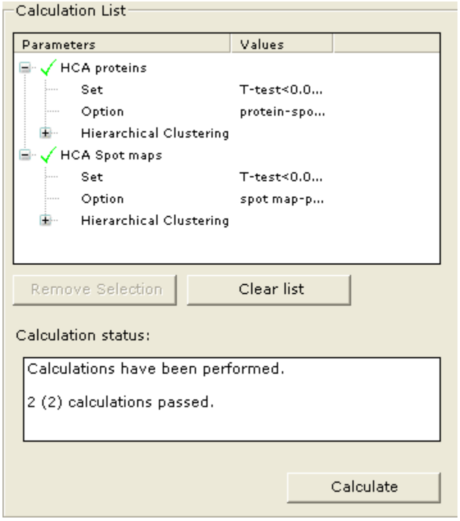
Calculation status:

Calculations pending

8 Click **Calculate** to start the calculation.

During calculation the status of each calculation is indicated by an icon in front of the calculation and the progress of the calculations are displayed by a progress bar.

9 When the calculations have been performed, the status of the calculations is displayed in the Calculation status area.



Calculation List

Parameters	Values
✓ HCA proteins	
Set	T-test<0.0...
Option	protein-spo...
Hierarchical Clustering	
✓ HCA Spot maps	
Set	T-test<0.0...
Option	spot map-p...
Hierarchical Clustering	

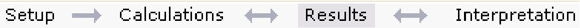
Calculation status:

Calculations have been performed.

2 (2) calculations passed.

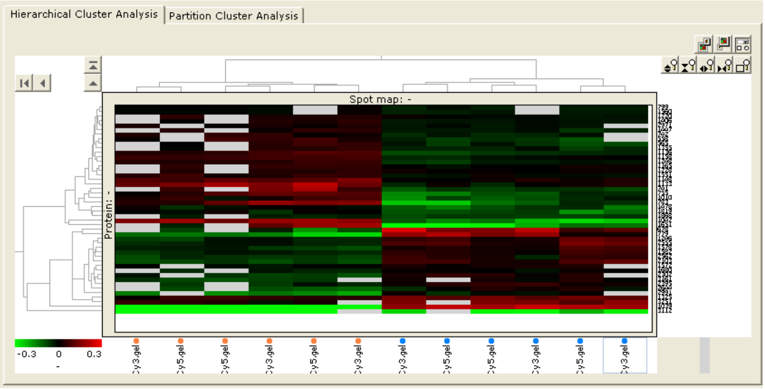
15.8.2 View the results of the hierarchical clustering

1 Select the **Results** step in the workflow area.



The **Results** window is displayed.

2 Select **Pattern Analysis** in the Results bar. The results for the hierarchical clustering are displayed in the Results view.



3 The following analyses of the results can be performed:

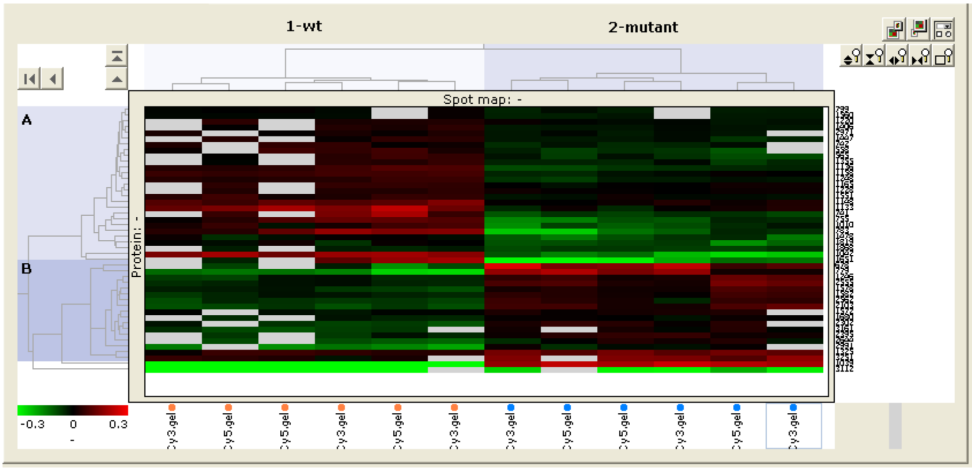
- *The spot maps in each experimental group have been clustered together*
  The spot maps have been divided into two clusters. This can be determined by looking at the dendrogram for the spot maps and the experimental group color coding of the spot maps in the two groups (blue and orange).

- *A rough estimation of the number of protein groups with the same expression patterns can be made*
  Approximately two main groups of proteins can be seen: A and B, although some proteins deviate from their group and some proteins are not included in the main groups.

  In group A, most proteins are up-regulated in the wt group and down-regulated in the mutant group.

  In group B, most proteins are down-regulated in the wt group and up-regulated in the mutant group.

Tip! *For information on how to zoom within the dendrogram and on how to select proteins and spot maps, see section 9.4, Analyze the results of the hierarchical clustering.*

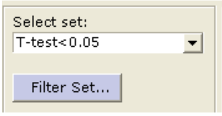


4 The analysis has confirmed the PCA results and further information on the protein expression patterns among the significantly differentially expressed proteins has been obtained.

15.9 Select and create sets with proteins from which to generate pick lists

An overview of the data in the form of PCA and hierarchical clustering has now been performed. A set will now be created with proteins for picking. This is performed by filtering the **T-test<0.05** set to remove all missing values.

To create the set with proteins for picking:



1 Still in the Pattern Analysis results view, make sure the **T-test<0.05** set is selected.

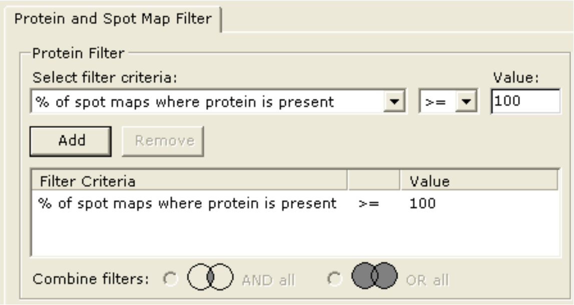
2 Click **Filter set...** . The **Filter** dialog opens.

3 To remove proteins with missing values and extract proteins with log standard abundance value difference of 0.1 among the spot maps, select the protein filter as follows:

a. Select the filter criteria **% of spot maps where protein is present**.

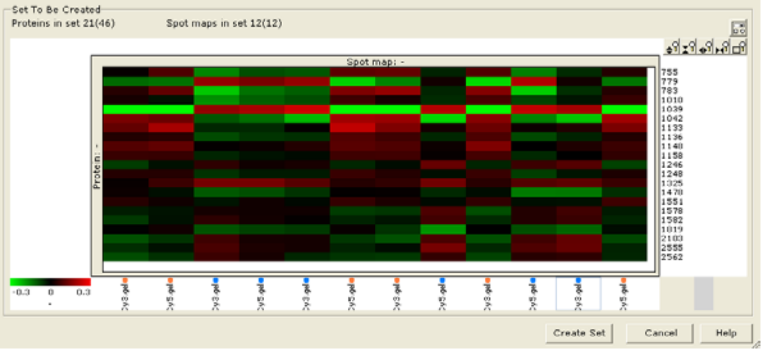
b. Choose the operator **>=** and enter the value **100**.

c. Click **Add**. The filter criteria is added to the list below.



Filter Criteria		Value
% of spot maps where protein is present	>=	100

4 Click **Apply FIlter** to apply the filter criteria to the base set. The heat map will be updated and information on the number of remaining spot maps and proteins are displayed in the **Set To Be Created** area.

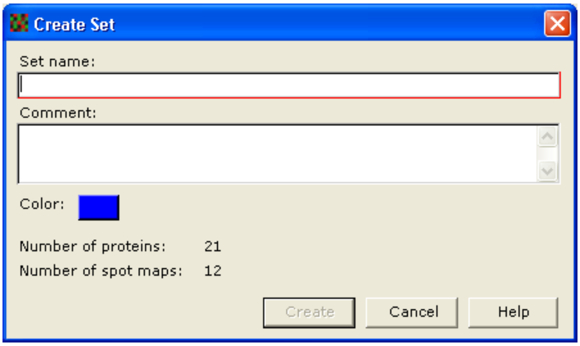


Set To Be Created

Proteins in set 21(46)

Spot maps in set 12(12)

5 Click **Create set**. The **Create set** dialog is displayed.



Create Set

Set name:

Comment:

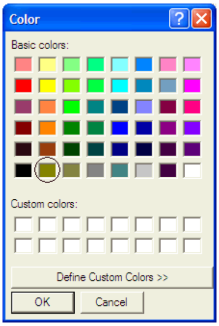
Color: 🟦

Number of proteins: 21

Number of spot maps: 12

6 Type in *"Tutorial I picking"* in the **Set name** field.

7 Click the colored button. The **Color** dialog opens. Select olive green and click **OK** to change color and return to the **Create set** dialog.

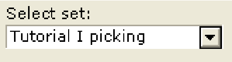


8 Click **Create**. The **Tutorial I picking** set is created. It will be added to the **Select set** list.

15.10Create a pick list

A set with proteins of interest has been created. A pick list will now be created and applied to the pick gel (for picking of gel spots) in BVA. MS analysis can then be performed. After performing the protein ID search in the appropriate database, obtained MS data (accession number of protein ID's) can be imported into EDA.

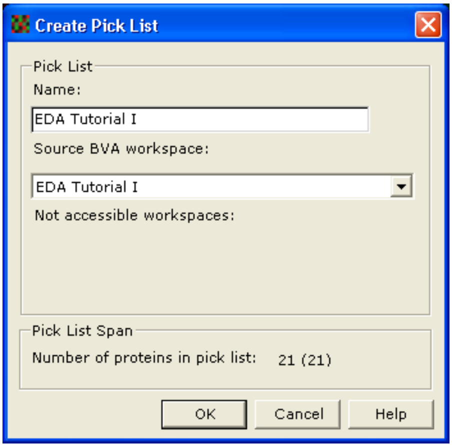
To create the pick list:



Select set:
Tutorial I picking

1 Select the **Tutorial I picking** set in the **Select set** pop-up dialog.

2 Select **Tools:Create Pick List in BVA...** in the menu bar. The **Create Pick List** dialog is displayed.



Create Pick List

Pick List

Name:
EDA Tutorial I

Source BVA workspace:
EDA Tutorial I

Not accessible workspaces:

Pick List Span

Number of proteins in pick list: 21 (21)

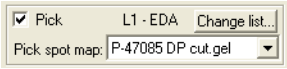
3 Type in *"EDA Tutorial I"* in the **Name** field.

4 As only one BVA workspace was imported into EDA, this workspace is selected in the **Source BVA workspace** field.

5 Click **OK** to create the pick list.

BVA opens with the Protein Table displayed and the proteins in the pick list are assigned to Pick.

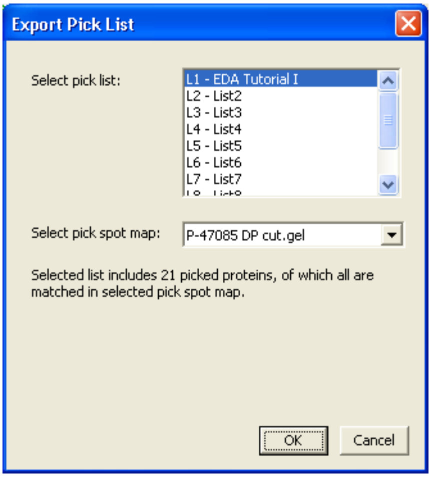
The created pick list name is shown below the Protein Table, and the pick gel is selected in the **Pick spot map** field.



☑ Pick L1 - EDA

Pick spot map: P-47085 DP cut.gel

6 In BVA, select **File:Export Pick List...**. The **Export Pick List** dialog opens with the correct pick list and pick spot map selected.



Selected list includes 21 picked proteins, of which all are matched in selected pick spot map.

7 Click **OK**. In the dialog that appears, choose a folder in which to save the pick list, type in the name *"EDA Tutorial I pick list"*, make sure the file format is *.txt and click **Save**. The pick list has been created.

Tip! *A pre-prepared pick list (EDA Tutorial I pick list finished.txt) is also available on the Tutorial DVD. It is possible to use the created pick list or the pick list on the DVD.*

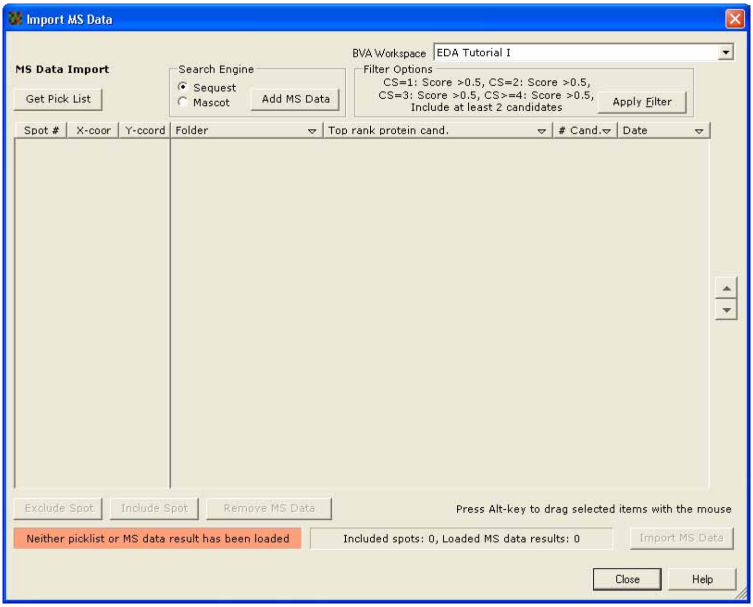
8 Close BVA without saving the BVA workspace.

15.11Import MS data

A pick list has been created and MS analysis has been performed. The MS data (protein ID data) will now be imported into EDA. The accession numbers from the MS data are needed to identify the proteins in the different databases.

To import the MS data:

1 Select **File:Import MS data...** in the menu bar. The **Import MS Data** dialog is displayed.



BVA Workspace: EDA Tutorial I

Search Engine: ● Sequest ○ Mascot

Filter Options: CS=1: Score >0.5, CS=2: Score >0.5, CS=3: Score >0.5, CS>=4: Score >0.5, Include at least 2 candidates

Press Alt-key to drag selected items with the mouse

Neither picklist or MS data result has been loaded

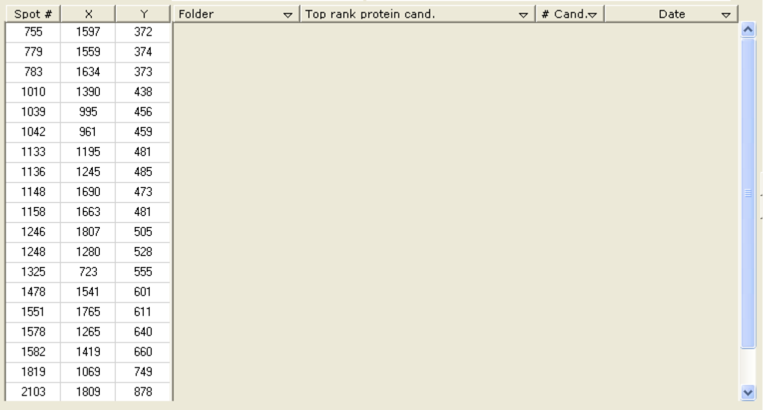
Included spots: 0, Loaded MS data results: 0

2 The BVA workspace included in EDA is displayed in the **BVA Workspace** field.

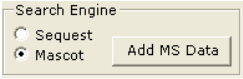
3 Click **Get Pick List** and locate and open the pick list you just created in the dialog that appears.

Alternatively, open the pick list *EDA Tutorial I pick list finished.txt* included on the Tutorials DVD.

The pick list is displayed in the left part of the dialog. The **Spot#** column shows the Master spot number of the protein and the **X-coord** and **Y-coord** columns show the coordinates on the pick gel.



Spot #	X	Y	Folder	Top rank protein cand.	# Cand.	Date
755	1597	372				
779	1559	374				
783	1634	373				
1010	1390	438				
1039	995	456				
1042	961	459				
1133	1195	481				
1136	1245	485				
1148	1690	473				
1158	1663	481				
1246	1807	505				
1248	1280	528				
1325	723	555				
1478	1541	601				
1551	1765	611				
1578	1265	640				
1582	1419	660				
1819	1069	749				
2103	1809	878				



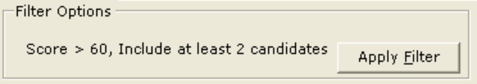
Search Engine

○ Sequest

● Mascot

4 Select the **Mascot** radio button in the **Search engine** area.

5 Use the default settings in the **Filter Options** area. Proteins with score > 60 or at least two candidates from each search will be imported.

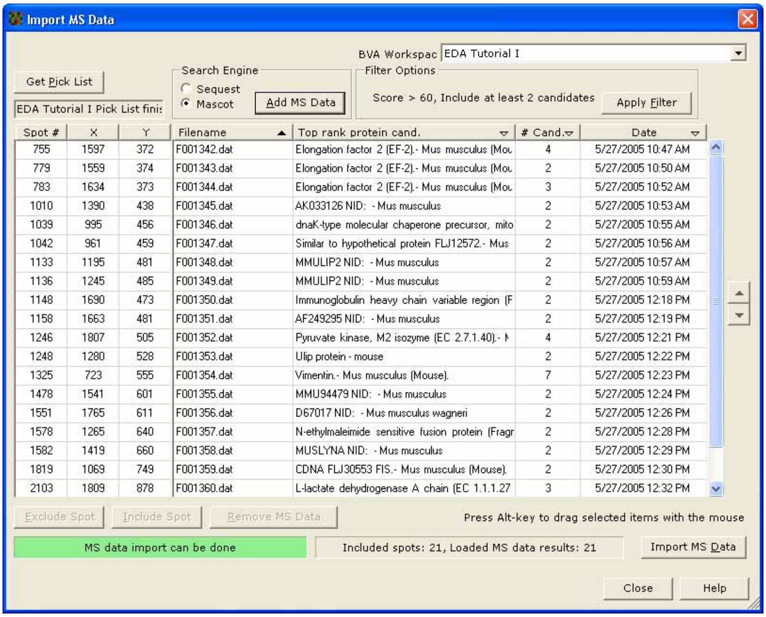


Filter Options

Score > 60, Include at least 2 candidates

6 Click **Add MS data** in the **Search engine** area. The **Select Mascot result file** dialog appears.

Locate the MS data files (one Mascot PMF (.dat) file/spot), select the files (using Shift + click or Ctrl + click) and click **Open**. The MS data is displayed in the right part of the **Import MS data** dialog. Only the top ranked protein is shown in the table.



BVA Workspac: EDA Tutorial I

EDA Tutorial I Pick List finis

Search Engine: ○ Sequest ● Mascot

Filter Options: Score > 60, Include at least 2 candidates

Spot #	X	Y	Filename	Top rank protein cand.	# Cand.	Date
755	1597	372	F001342.dat	Elongation factor 2 (EF-2).- Mus musculus (Mou	4	5/27/2005 10:47 AM
779	1559	374	F001343.dat	Elongation factor 2 (EF-2).- Mus musculus (Mou	2	5/27/2005 10:50 AM
783	1634	373	F001344.dat	Elongation factor 2 (EF-2).- Mus musculus (Mou	3	5/27/2005 10:52 AM
1010	1390	438	F001345.dat	AK033126 NID: - Mus musculus	2	5/27/2005 10:53 AM
1039	995	456	F001346.dat	dnaK-type molecular chaperone precursor, mito	2	5/27/2005 10:55 AM
1042	961	459	F001347.dat	Similar to hypothetical protein FLJ12572.- Mus	2	5/27/2005 10:56 AM
1133	1195	481	F001348.dat	MMULIP2 NID: - Mus musculus	2	5/27/2005 10:57 AM
1136	1245	485	F001349.dat	MMULIP2 NID: - Mus musculus	2	5/27/2005 10:59 AM
1148	1690	473	F001350.dat	Immunoglobulin heavy chain variable region (F	2	5/27/2005 12:18 PM
1158	1663	481	F001351.dat	AF249295 NID: - Mus musculus	2	5/27/2005 12:19 PM
1246	1807	505	F001352.dat	Pyruvate kinase, M2 isozyme (EC 2.7.1.40).- M	4	5/27/2005 12:21 PM
1248	1280	528	F001353.dat	Ulip protein - mouse	2	5/27/2005 12:22 PM
1325	723	555	F001354.dat	Vimentin.- Mus musculus (Mouse).	7	5/27/2005 12:23 PM
1478	1541	601	F001355.dat	MMU94479 NID: - Mus musculus	2	5/27/2005 12:24 PM
1551	1765	611	F001356.dat	D67017 NID: - Mus musculus wagneri	2	5/27/2005 12:26 PM
1578	1265	640	F001357.dat	N-ethylmaleimide sensitive fusion protein (Fragr	2	5/27/2005 12:28 PM
1582	1419	660	F001358.dat	MUSLYNA NID: - Mus musculus	2	5/27/2005 12:29 PM
1819	1069	749	F001359.dat	CDNA FLJ30553 FIS.- Mus musculus (Mouse).	2	5/27/2005 12:30 PM
2103	1809	878	F001360.dat	L-lactate dehydrogenase A chain (EC 1.1.1.27	3	5/27/2005 12:32 PM

Press Alt-key to drag selected items with the mouse

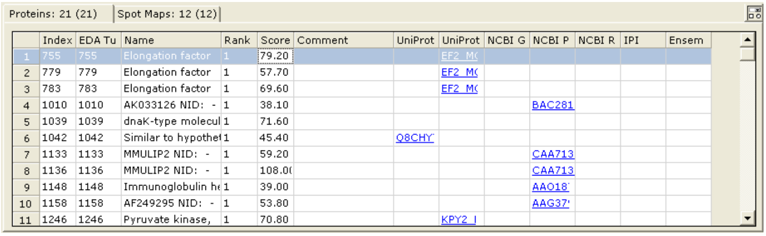
MS data import can be done

Included spots: 21, Loaded MS data results: 21

7 As the MS analysis of the picked proteins has been run in the same order as in the pick list, the MS data is correctly matched to the pick list (each row in the table contains the protein to pick and the corresponding MS data for that protein).

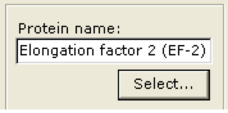
8 Click **Import MS data** to import it into EDA. Click **Close** to the **Import MS data** dialog.

9 The MS data will be displayed in the Protein table.



Proteins: 21 (21) | Spot Maps: 12 (12)

	Index	EDA Tu	Name	Rank	Score	Comment	UniProt	UniProt	NCBI G	NCBI P	NCBI R	IPI	Ensem
1	755	755	Elongation factor	1	79.20			EF2_MC					
2	779	779	Elongation factor	1	57.70			EF2_MC					
3	783	783	Elongation factor	1	69.60			EF2_MC					
4	1010	1010	AK033126 NID: -	1	38.10					BAC281			
5	1039	1039	dnaK-type molecul	1	71.60								
6	1042	1042	Similar to hypothet	1	45.40		Q8CHY						
7	1133	1133	MMULIP2 NID: -	1	59.20					CAA713			
8	1136	1136	MMULIP2 NID: -	1	108.0(					CAA713			
9	1148	1148	Immunoglobulin he	1	39.00					AAO18			
10	1158	1158	AF249295 NID: -	1	53.80					AAG37'			
11	1246	1246	Pyruvate kinase,	1	70.80			KPY2_I					

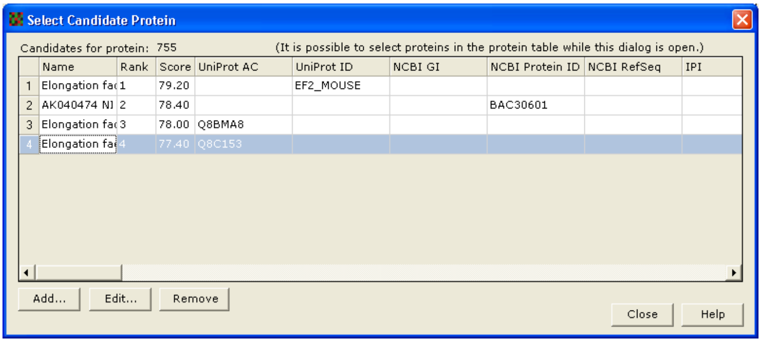


Protein name:

Elongation factor 2 (EF-2)

10 Select a protein candidate in the Protein table and click **Select...** to view details for the imported protein candidate.

The **Select Candidate Protein** dialog is displayed.



Select Candidate Protein

Candidates for protein: 755 (It is possible to select proteins in the protein table while this dialog is open.)

	Name	Rank	Score	UniProt AC	UniProt ID	NCBI GI	NCBI Protein ID	NCBI RefSeq	IPI
1	Elongation fac	1	79.20		EF2_MOUSE				
2	AK040474 NI	2	78.40				BAC30601		
3	Elongation fac	3	78.00	Q8BMA8					
4	Elongation fac	4	77.40	Q8C153					

In this dialog it is possible to select another candidate protein to be displayed in the Protein table.

11 Web links for the different proteins will appear in the Protein table. Click a link to get information about that protein.

16 Tutorial II - Classification of ovarian cancer biopsies

16.1 Objective

This tutorial describes how to perform differential expression analysis and PCA on a data set. It also describes how to find markers among the differentially expressed proteins, how to create a classifier using the markers and how to classify new samples using the classifier.

It also covers the central concept of how to create sets.

The purpose of this tutorial is to teach how to:

- Create a base set manually
- Perform differential expression analysis (One-way ANOVA) in EDA
- Perform discriminant analysis

16.2 Experiment overview

16.2.1 Introduction

A study of human ovarian cancer, in which 24 patients are included, was performed. Biopsy material from all patients was classified by pathologists into one of the following groups: normal, benign and malignant. All patient samples were run in duplicate. A subgroup of 5 patients of “unknown” class are to be classified using a learning algorithm and the result compared to the pathologists biopsy classification.

The aim of the experiment is to:

- identify biomarkers that can discriminate between the normal, benign and malignant classes (by using the biopsy material from the patients that was classified by pathologists)
- create a classifier and classify the patients in the “unknown” group (and compare the results with the results from the classification by pathologists)

16.2.2 Experimental design

Table 16-1 gives an overview of the experimental design in EDA with experimental groups, colors for the experimental groups and number of spot maps.

Experimental group	Color	# spot maps
Benign	Red	8
Malignant	Green	9
Normal	Blue	7
Unknown	Yellow	5 of which 2 are benign, 1 is malignant and 2 are normal according to the classification by pathologists.

Table 16-1. Experimental design.

Biopsy material was taken from the patients and divided into two sets:

- **Biopsies** set
  Patients that were classified by pathologists (normal, benign or malignant) (experimental groups: **Normal**, **Benign** and **Malignant**).

- **Unknowns** set
  Patients that are going to be classified using discriminant analysis in EDA (experimental group: **Unknown**) and then compared to the result from the classification by pathologists.

16.2.3 Basic work already performed

- Pre-processing of the gels in DIA and the BVA module have been performed, giving three BVA workspaces (Normal, Benign and Malignant).

- The BVA workspaces have been imported into EDA and are linked by a common Master. Five spot maps have been placed in a new group, Unknown, that will be classified by EDA.

16.3 Workflow overview

- Copy the tutorial file to your own project
- Start EDA
- *Set up the EDA workspace:* Open the created EDA workspace, edit the colors of the groups and create the base set manually
- *Create two sets:* **Biopsies** and **Unknown**
- Perform differential expression analysis on the **Biopsies** set
- Create a new set with filtered results
- Perform PCA calculations on the new set
- Perform discriminant analysis

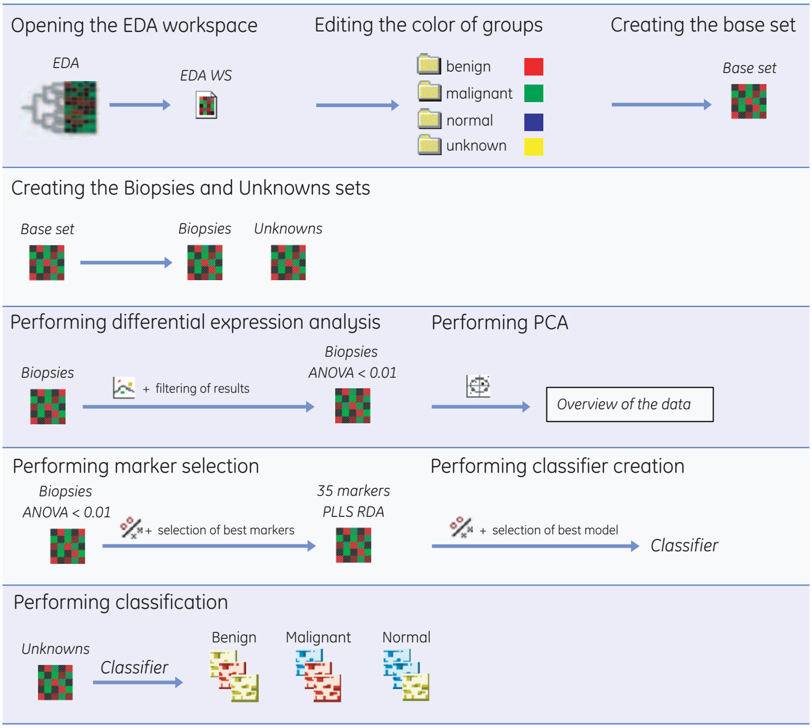

Opening the EDA workspace
EDA
EDA WS
Editing the color of groups
benign
malignant
normal
unknown
Creating the base set
Base set
Creating the Biopsies and Unknowns sets
Base set
Biopsies
Unknowns
Performing differential expression analysis
Biopsies
+ filtering of results
Biopsies
ANOVA < 0.01
Performing PCA
Overview of the data
Performing marker selection
Biopsies
ANOVA < 0.01
+ selection of best markers
35 markers
PLLS RDA
Performing classifier creation
+ selection of best model
Classifier
Performing classification
Unknowns
Classifier
Benign
Malignant
Normal

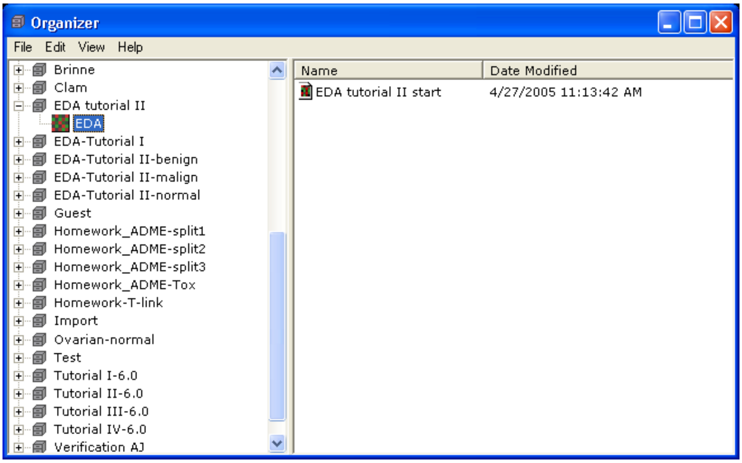
Fig 16-1. Workflow overview in EDA.

16.4 Copy the tutorial file to your own project

Before starting to work with this tutorial, copy the EDA Tutorial II workspace into your own personal project as follows:

1 In the DeCyder 2D start screen, click the **Organizer** button. The **Organizer** opens.

2 Double-click the **EDA tutorial II** project and click the EDA icon. The **EDA tutorial II start** workspace is displayed to the right.



Name	Date Modified
EDA tutorial II start	4/27/2005 11:13:42 AM

3 Right-click on the **EDA tutorial II start** workspace and select **Copy**.

4 Double-click the project with your personal files and click the EDA icon.

Alternatively, select **File:New project** to create a new project in the database in which to save your personal work on tutorial files. The **Create new project** dialog is displayed.

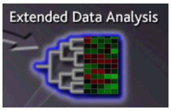
5 Enter a name for the project and click **OK** to create the project and return to the **Organizer**.

6 Select the created project.

7 In the right panel of the **Organizer**, right-click and select **Paste**. The **EDA tutorial II start** workspace is copied into the project.

16.5 Start EDA

1 Start DeCyder 2D Software, see section 2.3.

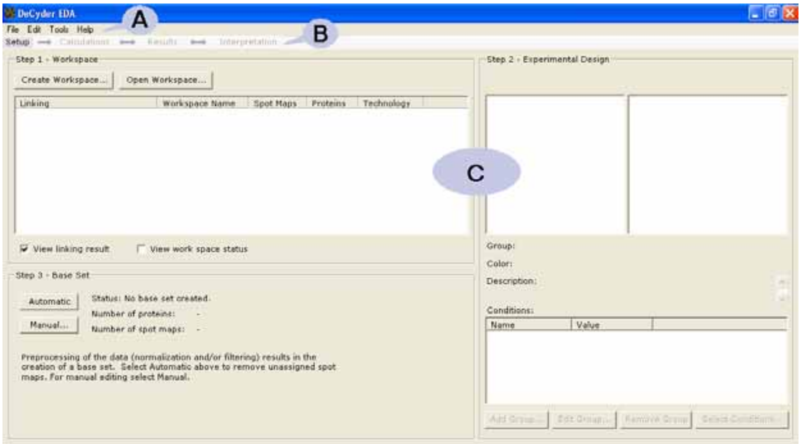


2 Click the Extended Data Analysis (EDA) icon in the DeCyder 2D main window.

3 EDA will open displaying the DeCyder EDA main screen, which is divided into three areas:

- menu bar (A)
- workflow area (B)
- work area (C)

Depending on the currently selected step in the workflow area, the work area will appear different. In the beginning, the first step in the workflow, **Setup**, is selected and the **Setup** window is displayed in the work area.



Status: No base set created.
Number of proteins: -
Number of spot maps: -

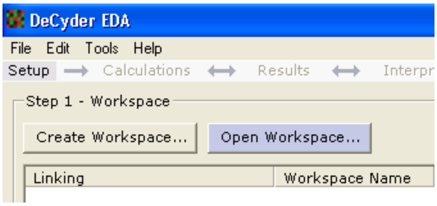
Preprocessing of the data (normalization and/or filtering) results in the creation of a base set. Select Automatic above to remove unassigned spot maps. For manual editing select Manual.

16.6 Set up the EDA workspace

An EDA workspace with the correct experimental design has been created. Setting up the workspace includes opening the EDA workspace and creating the base set.

16.6.1 Open the EDA Workspace

1 Click **Open Workspace...** in the **Step 1 - Workspace area** of the **Setup** window.

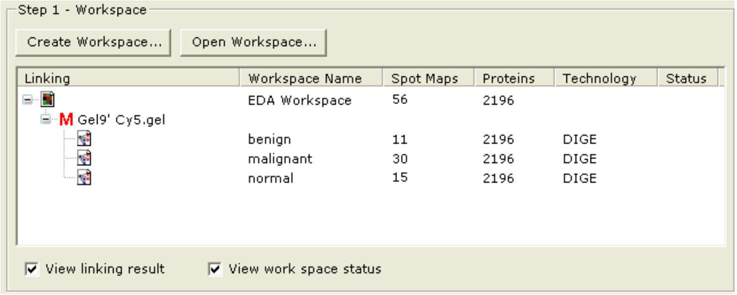


The **Open EDA Workspace dialog**is displayed.

2 Select the appropriate project (in the left panel) and click the EDA icon. The EDA workspaces in that project are displayed to the right.

3 Select **EDA tutorial II start** (in the right panel) and click **Open**.

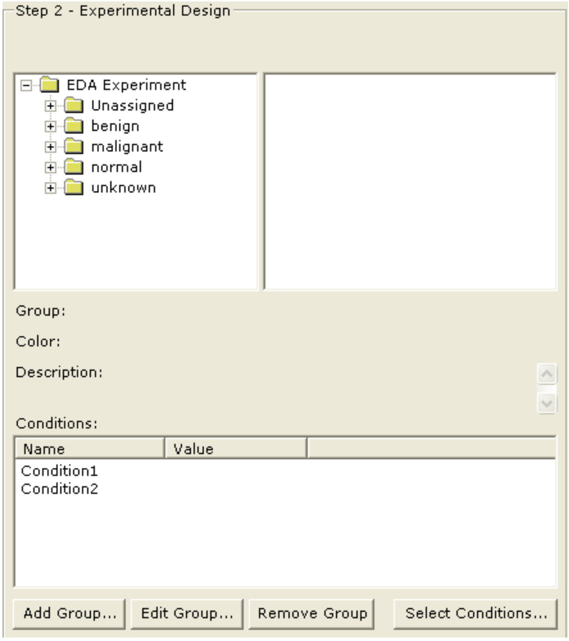
4 The EDA workspace is displayed in the **Step 1 - Workspace** area of the **Setup** window.



Linking	Workspace Name	Spot Maps	Proteins	Technology	Status
	EDA Workspace	56	2196		
M Gel9' Cy5.gel					
	benign	11	2196	DIGE	
	malignant	30	2196	DIGE	
	normal	15	2196	DIGE	

☑ View linking result ☑ View work space status

5 The experimental design for the EDA workspace in displayed in **Step 2 - Experimental Design** area of the **Setup** window. The colors of the groups need to be edited to facilitate the visualization of results later on, see section 16.6.2.



Group:

Color:

Description:

Conditions:

Name	Value
Condition1	
Condition2	

16.6.2 Edit the colors of the groups

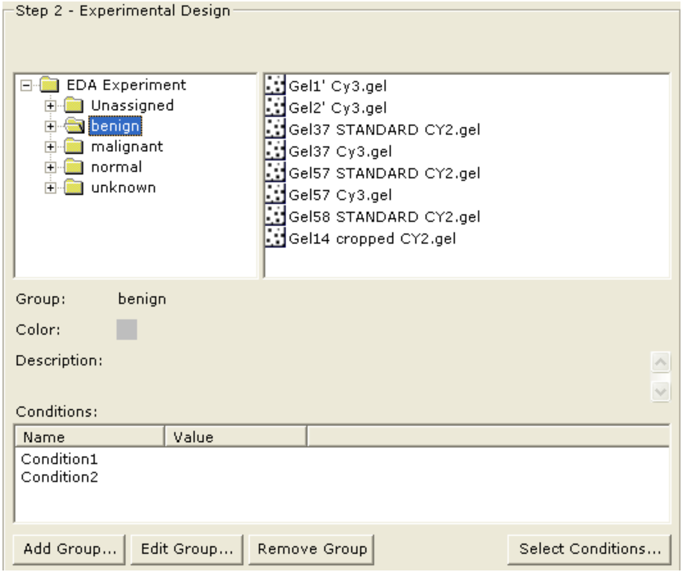
Different colors for the experimental groups facilitates the analyses in EDA. The different groups in this workspace have the same color, so the color needs to be changed according to the table below.

Experimental group	Change color to:
Benign	Red
Malignant	Green
Normal	Blue
Unknown	Yellow

Table 16-2. Colors for the experimental groups.

Change the color of the groups as follows:

1 Select the first group, **benign**, for which to edit the color. The name and color of the group is displayed in the **Group** and **Color** fields.



Step 2 - Experimental Design

- EDA Experiment
  - Unassigned
  - benign
  - malignant
  - normal
  - unknown

Gel1' Cy3.gel
Gel2' Cy3.gel
Gel37 STANDARD CY2.gel
Gel37 Cy3.gel
Gel57 STANDARD CY2.gel
Gel57 Cy3.gel
Gel58 STANDARD CY2.gel
Gel14 cropped CY2.gel

Group: benign

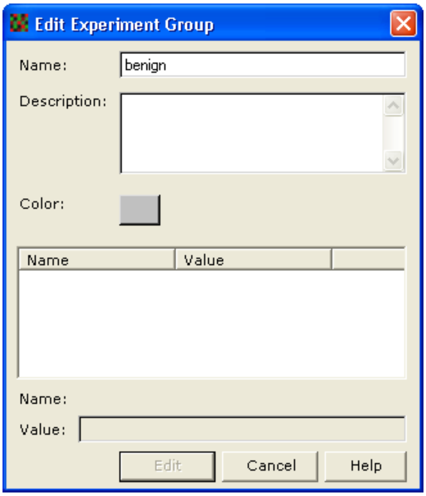
Color: ▨

Description:

Conditions:

Name	Value
Condition1	
Condition2	

2 Click **Edit Group...** . The **Edit Experiment Group** dialog is displayed.



Edit Experiment Group

Name: benign

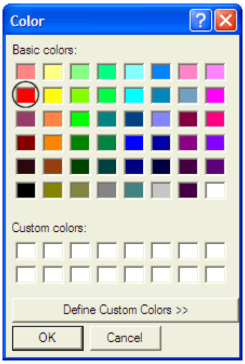
Description:

Color:

Name	Value	

Name:

Value:



3 Click the colored button in the **Color** field to open the **Color** dialog.

4 Select red and click **OK**.

5 Click **Edit**. The color is edited for the group.

6 Repeat steps 1-5 for the other groups and select colors according to Table 16-2.

16.6.3 Create the base set

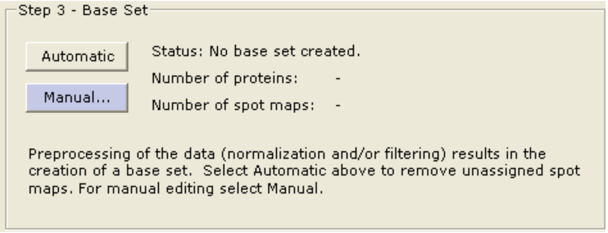
A base set must always be created before any analysis can be performed. When the base set has been created, the rest of the steps in the workflow area become activated and new sets can be created and calculations can be performed.

The base set can be created either manually or automatically. In this tutorial, the base set is created manually. The reason for this is that the workspace contains missing values (that need to be removed from the data set) and this is not performed if the base set is created automatically.

Note: *Missing values arise when a spot is not present in one or several spot maps. If the spot is absent in many spot maps, this will affect the PCA calculation in a negative way. Therefore, spots with many missing values should be removed from the set.*

Create the base set as follows:

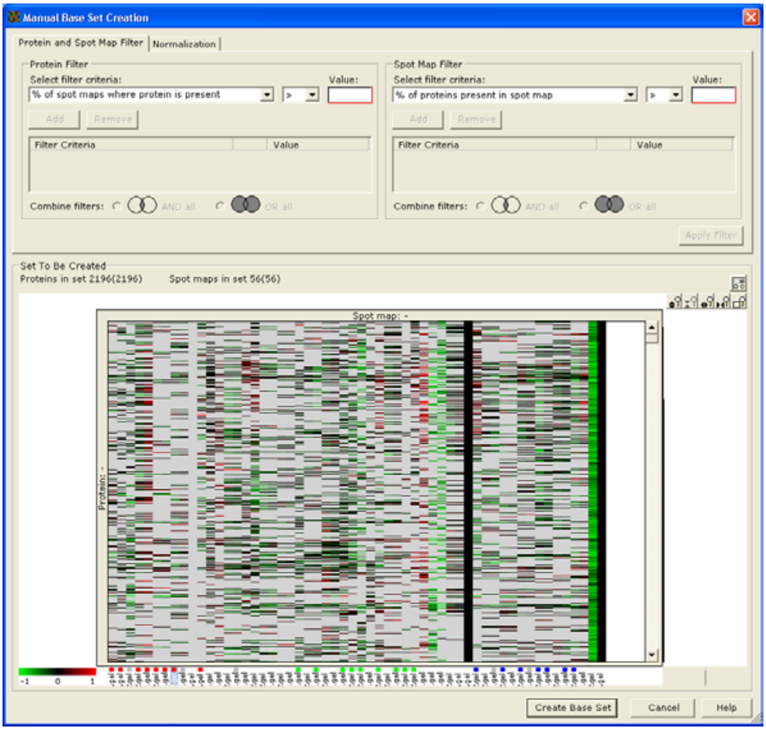
1 Click **Manual...** in the **Step 3 - Base Set Creation** area.



Status: No base set created.
Number of proteins: -
Number of spot maps: -

Preprocessing of the data (normalization and/or filtering) results in the creation of a base set. Select Automatic above to remove unassigned spot maps. For manual editing select Manual.

The **Manual Base Set Creation dialog** opens displaying the **Protein and Spot Map Filter** tab by default.



Protein Filter: Select filter criteria: % of spot maps where protein is present >

Spot Map Filter: Select filter criteria: % of proteins present in spot map >

Set To Be Created

Proteins in set 2196(2196)

Spot maps in set 56(56)

The data set is displayed in the form of a heat map in the **Set To Be Created** area. This area also displays the number of proteins and spot maps currently included in the base set to be created.

For more information about the heat map, how to change settings and how to zoom within the heat map, see section 3.3.

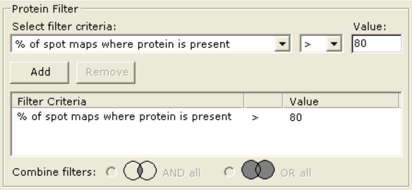
2 To remove all spots that are not present in at least 80% of the spot maps (missing in 20% of the spot maps), add the following filter criteria to the list:

a. Select the filter criteria **% of spot maps where protein is present**.

b. Choose the operator **>**.

c. Enter the value **80**.

d. Click **Add**. The filter criteria is added to the list below.



Protein Filter

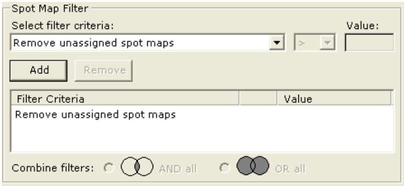
Select filter criteria: % of spot maps where protein is present > Value: 80

Filter Criteria		Value
% of spot maps where protein is present	>	80

Combine filters: ○ AND all ○ OR all

3 *To remove all unassigned spot maps (spot maps contained within the* ***Unassigned*** *group in the experimental design should not be included in calculations):*

Select the spot map filter **Remove unassigned spot maps** and click **Add** to add it to the filter criteria list.



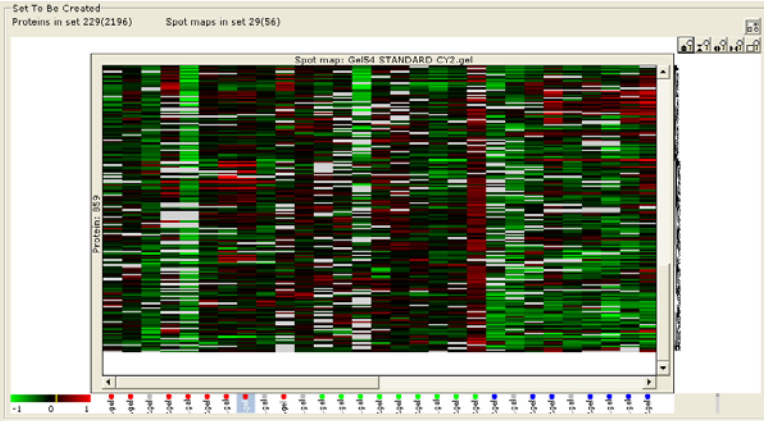
Spot Map Filter

Select filter criteria: Remove unassigned spot maps Value:

Filter Criteria	Value
Remove unassigned spot maps	

Combine filters: ○ AND all ○ OR all

4 Click **Apply Filter** to apply the two filter criteria to the base set. The heat map will be updated and information on the number of remaining spot maps and proteins will be displayed in the **Set To Be Created** area.



Set To Be Created

Proteins in set 229(2196) Spot maps in set 29(56)

5 Click **Create Base Set**. During the creation, a dialog showing the progress is displayed.

6 When the base set has been created, the status **Base set created**- **Analysis is now possible** is displayed in the **Status** field. The number of proteins and spot maps included in the base set are also displayed.

7 Select **File:Save** to save the changes in the workspace.

Tip! *It is recommended to save the workspace regularly.*

16.7 Create sets on which to perform calculations

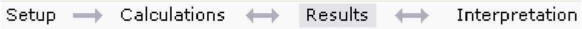
The EDA workspace has been set up. Two sets are now going to be created from the base set, that will be used in further analyses:

- **Biopsies** set
  This set will contain the spot maps and proteins from the **Normal**, **Benign** and **Malignant** experimental groups. This set will be used to create the classifier. All calculations, except classification, will be performed on this set.
- **Unknowns** set
  This set will contain the spot maps and proteins from the **Unknown** experimental group. The spot maps in this set will be classified once a classifier has been created.

16.7.1 Create the Unknowns set

The two sets will be created by selecting data manually.

1 Select the **Results** step in the workflow area.

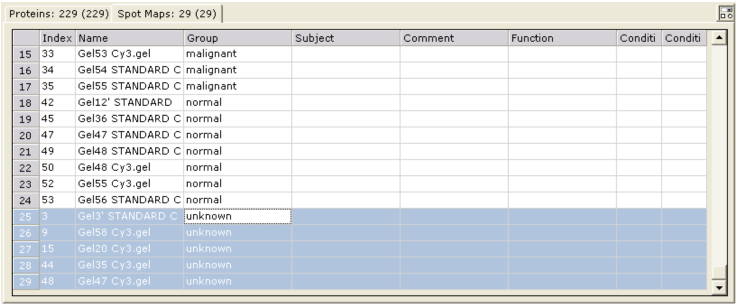


2 In the Spot Map table (in the **Results** window), select the proteins and spot maps that belong to the group **Unknown** (that will be included in the **Unknowns** set to be created) as follows:

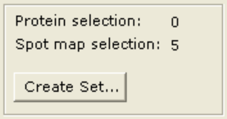
a. Click the **Spot Maps** tab to display the Spot Map table.

b. Click the column header **Group** to sort the table according to experimental group name (**Benign**, **Malignant**, **Normal** and **Unknown**).

c. Select all spot maps in the **Unknown** group by clicking the first cell with an unknown spot map, pressing the **Shift** button and clicking the last cell containing an unknown spot map. All spot maps in the **Unknown** group are now selected.



	Index	Name	Group	Subject	Comment	Function	Conditi	Conditi
15	33	Gel53 Cy3.gel	malignant					
16	34	Gel54 STANDARD C	malignant					
17	35	Gel55 STANDARD C	malignant					
18	42	Gel12' STANDARD	normal					
19	45	Gel36 STANDARD C	normal					
20	47	Gel47 STANDARD C	normal					
21	49	Gel48 STANDARD C	normal					
22	50	Gel48 Cy3.gel	normal					
23	52	Gel55 Cy3.gel	normal					
24	53	Gel56 STANDARD C	normal					
25	3	Gel3' STANDARD C	unknown					
26	9	Gel58 Cy3.gel	unknown					
27	15	Gel20 Cy3.gel	unknown					
28	44	Gel35 Cy3.gel	unknown					
29	48	Gel47 Cy3.gel	unknown					

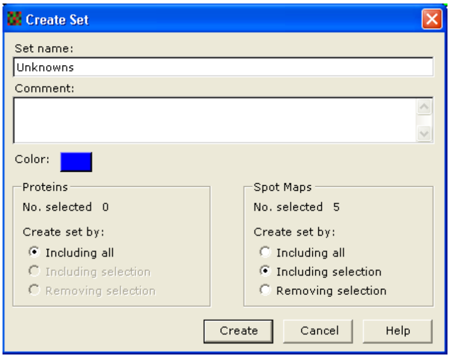

Protein selection: 0
Spot map selection: 5

Tip! *In the Set area it is always possible to see how many proteins and spot maps are currently selected.*

3 Click **Create Set...** in the Set area. The **Create set** dialog is displayed.

4 Enter "*Unknowns*" in the **Set name** field.

5 Make sure the **Including all** radio button is selected in the **Proteins** area and the **Including selection** radio button is selected in the **Spot Maps** area.



Set name: Unknowns

Comment:

Color: 🟦

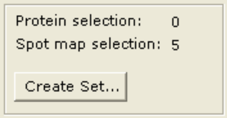
Proteins
No. selected 0
Create set by:
● Including all
○ Including selection
○ Removing selection

Spot Maps
No. selected 5
Create set by:
○ Including all
● Including selection
○ Removing selection

6 This means that the proteins and spot maps in the **Unknown** group will be included in the set to be created.

7 Click **Create**. The **Unknowns** set is created. It will be displayed in the **Select set** field.

16.7.2 Create the Biopsies set



Protein selection: 0

Spot map selection: 5

1 Keep the current selection of proteins and spot maps that were used to create the **Biopsies** set.

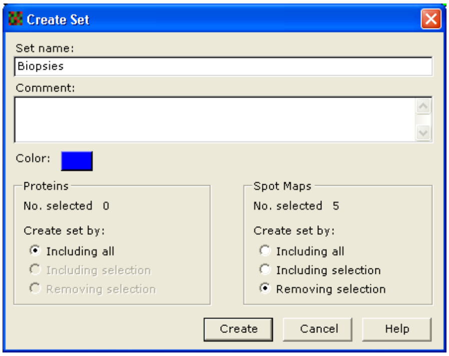
2 Click **Create Set...** in the Set area. The **Create set** dialog is displayed.

3 Enter "*Biopsies*" in the **Set name** field.

4 Make sure the **Including all** radio button is selected in the **Proteins** area.

5 Choose the **Removing selection** radio button in the **Spot Maps** area.

6 Click **Create**. The **Biopsies** set is created and will be displayed in the **Selected set** field.



Create Set

Set name: Biopsies

Comment:

Color: 🟦

Proteins

No. selected 0

Create set by:

● Including all

○ Including selection

○ Removing selection

Spot Maps

No. selected 5

Create set by:

○ Including all

○ Including selection

● Removing selection

16.8 Perform differential expression analysis

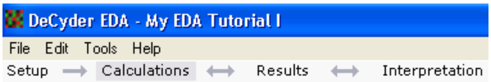
The two sets **Biopsies** and **Unknowns** have now been created.

Differential expression analysis will now be performed on the **Biopsies** set to find significantly differentially expressed proteins when comparing the three groups of normal, benign and malignant biopsy samples.

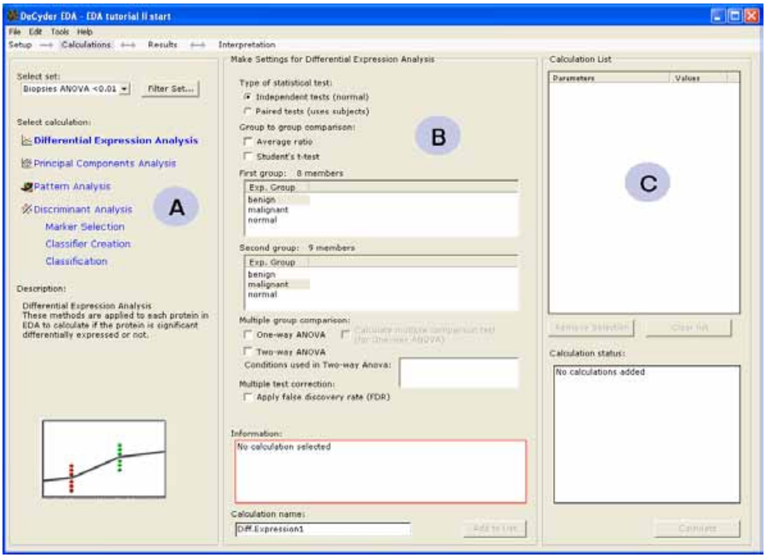
Note: *No calculations are performed on the Unknowns set because this set contains spot maps with "unknown" class, i.e. spot maps from the different classes. This set is only used for classification (see section 16.7).*

16.8.1 Set up the differential expression analysis calculation

1 Click **Calculations** in the workflow area.


DeCyder EDA - My EDA Tutorial I

The **Calculations window** opens displaying the settings for differential expression analysis by default.


DeCyder EDA - EDA tutorial II start

Select set: Biopsies ANOVA <0.01

Select calculation: Differential Expression Analysis

Description:
Differential Expression Analysis
These methods are applied to each protein in EDA to calculate if the protein is significant differentially expressed or not.

Make Settings for Differential Expression Analysis

Type of statistical test:
● Independent tests (normal)
○ Paired tests (uses subjects)

Group to group comparison:
□ Average ratio
□ Student's t-test

First group: 8 members

Exp. Group
benign
malignant
normal

Second group: 9 members

Exp. Group
benign
malignant
normal

Multiple group comparison:
□ One-way ANOVA □ Calculate multiple comparison test (for One-way ANOVA)
□ Two-way ANOVA
Conditions used in Two-way Anova:

Multiple test correction:
□ Apply false discovery rate (FDR)

Information:
No calculation selected

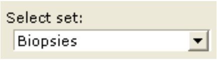
Calculation name: Diff.Expression1

Calculation List

Parameters	Values

Calculation status:
No calculations added

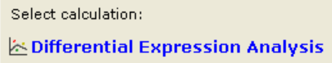
Fig 16-2. Calculations window. The window is divided into three areas: **Calculations**, **A** (where the set on which to perform calculations and the type of calculation are selected), **Select settings**, **B** (where settings for the selected calculation in **A** are entered) and **Calculation List**, **C** (where the added calculations are listed and can be calculated).



Select set: Biopsies

2 Select the **Biopsies** set in the **Select set** field.

3 By default, the **Differential Expression Analysis** is selected in the **Select calculation** area and the settings for the calculation displayed to the right.



Select calculation:

Differential Expression Analysis

4 Enter the following settings:

a. Choose **Independent tests (normal)** in the **Type of statistical tests** area.

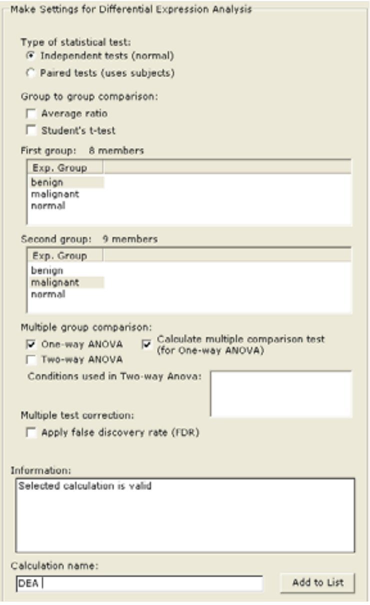
b. Check the **One-way ANOVA** box in the **Group to group comparison** area.

c. Check the **Calculate multiple comparison test (for One-Way ANOVA)**.

d. Leave all other boxes unchecked.

e. Type in *"DEA"* in the **Calculation name** field.

5 Click **Add to List** to add the calculation to the **Calculation List** to the right.



Make Settings for Differential Expression Analysis

Type of statistical test:
● Independent tests (normal)
○ Paired tests (uses subjects)

Group to group comparison:
□ Average ratio
□ Student's t-test

First group: 8 members

Exp. Group
benign
malignant
normal

Second group: 9 members

Exp. Group
benign
malignant
normal

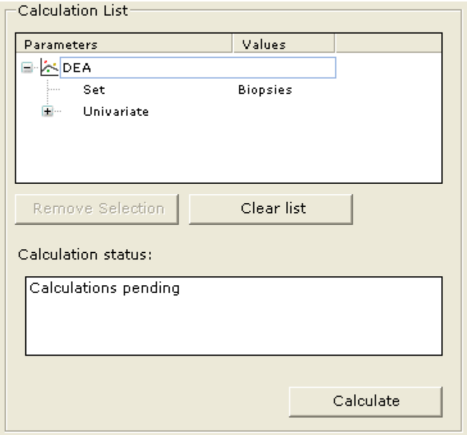
Multiple group comparison:
☑ One-way ANOVA
□ Two-way ANOVA
☑ Calculate multiple comparison test (for One-way ANOVA)

Conditions used in Two-way Anova:

Multiple test correction:
□ Apply false discovery rate (FDR)

Information:
Selected calculation is valid

Calculation name: DEA



Calculation List

Parameters	Values
DEA	
Set	Biopsies
Univariate	

Calculation status:
Calculations pending

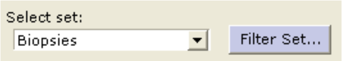
6 Click **Calculate** to start the calculation.
During calculation the status of the calculation is indicated by an icon in front of the calculation and the progress of the calculation is displayed by a progress bar.

16.8.2 Create new sets by filtering the results

The differential expression analysis calculation on the **Biopsies** set has been performed. The result of the calculation is now going to be filtered (extracting significantly differentially expressed proteins) and a new set created, containing the significantly differentially expressed proteins. The filter will extract proteins with an ANOVA p-value < 0.01.

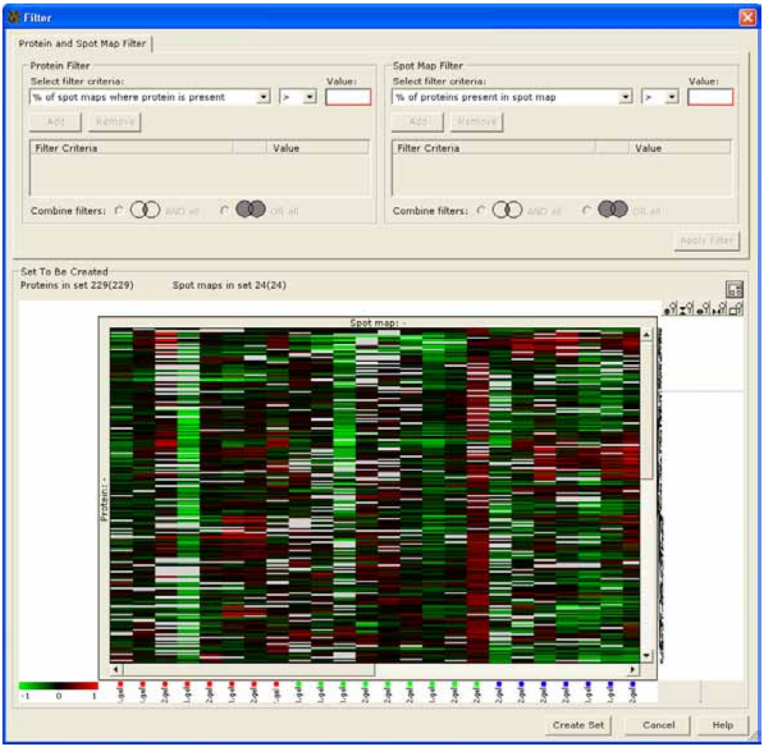
To filter the result and create a new set:

1 Still in the **Calculations** step, make sure the **Biopsies** set is selected in the **Select set** field and click **Filter Set**.



Select set: Biopsies

The **Filter dialog** is displayed.



Protein Filter
Select filter criteria: % of spot maps where protein is present >
Value:
Filter Criteria | Value
Combine filters: ○ AND all ○ OR all

Spot Map Filter
Select filter criteria: % of proteins present in spot map >
Value:
Filter Criteria | Value
Combine filters: ○ AND all ○ OR all

Set To Be Created
Proteins in set 229(229)   Spot maps in set 24(24)

Fig 16-3. Filter dialog.

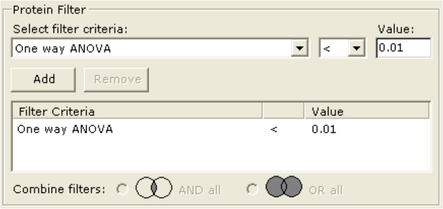
2 To extract all proteins with a p-value < 0.01 (from the ANOVA calculation), select the protein filter as follows:

a. Select the filter criteria **One-way ANOVA**.

b. Choose the operator **<**.

c. Enter the value **0.01**.

d. Click **Add**. The filter criteria is added to the list below.

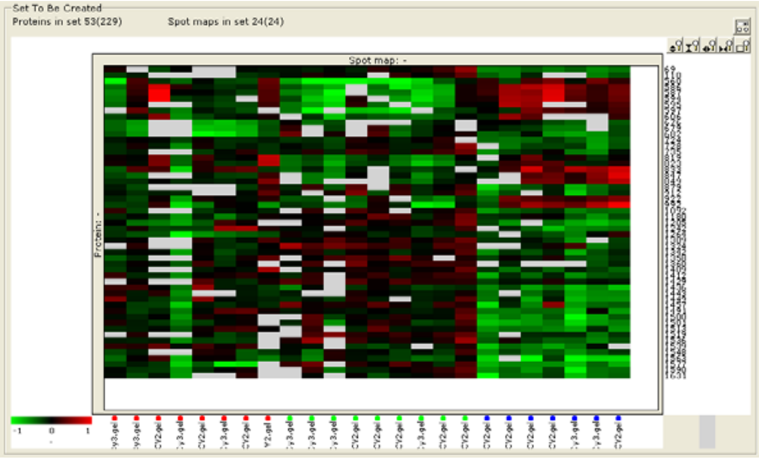


Protein Filter

Select filter criteria: One way ANOVA < Value: 0.01

Filter Criteria		Value
One way ANOVA	<	0.01

3 Click **Apply Filter** to apply the filter criteria to the **Biopsies** set. The heat map will be updated and information on the number of remaining spot maps and proteins displayed in the **Set To Be Created** area.



Set To Be Created

Proteins in set 53(229) Spot maps in set 24(24)

4 Click **Create Set**. The **Create set** dialog is displayed.

5 Enter *"Biopsies ANOVA < 0.01"* in the **Set name** field and enter *"Proteins with an ANOVA p-value < 0.01 included"* in the **Comment** field.

6 Click **Create**. The **Biopsies ANOVA < 0.01** set is created. Select this set in the **Select set** field to carry on with calculations on the new set

16.9 Perform PCA

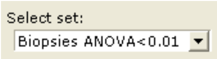
One set with significantly differentially expressed proteins has been created (**Biopsies ANOVA < 0.01**). PCA on this set will now be performed.

In the PCA, it is possible to see:

- Which proteins lie outside of a 95% significance level in expression (and should therefore be checked)
- The relation between proteins and spot maps.

PCA is mainly performed to obtain an overview of the data and to check that the data looks OK (e.g. there are no spot map and/or protein outliers).

16.9.1 Set up the PCA calculation



Select set:
Biopsies ANOVA<0.01

1 In the **Calculations** step, select the **Biopsies ANOVA < 0.01** set in the **Select set** field.

2 Select **Principal Component Analysis** in the **Select calculation** area. The settings for the calculation are displayed to the right.

Principal Components Analysis

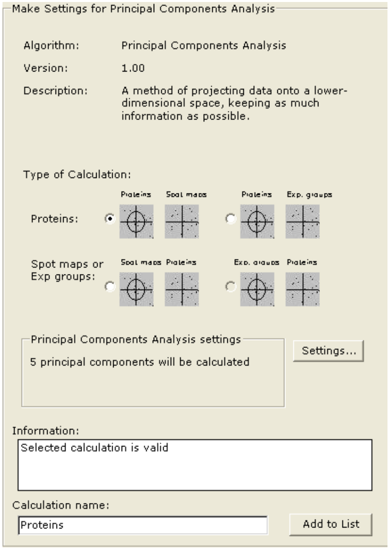
3 Enter the following settings for the PCA calculation:

a. In the **Type of Calculation** area, choose the left radio button in the **Proteins** area.

This setting will give an overview of the data, with proteins in the score plot (left plot) and spot maps in the loading plot (right plot).

b. Use the default settings displayed in the **Principal Component Analysis settings** area.

c. Type in *"Proteins"* in the **Calculation name** field.



Make Settings for Principal Components Analysis

Algorithm: Principal Components Analysis

Version: 1.00

Description: A method of projecting data onto a lower-dimensional space, keeping as much information as possible.

Type of Calculation:

Proteins: ● ○

Spot maps or Exp groups: ○ ○

Principal Components Analysis settings

5 principal components will be calculated

Information:

Selected calculation is valid

Calculation name: Proteins

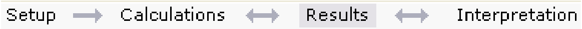
4 Click **Add to List** to add the calculation to the **Calculation List** to the right.

5 Click **Calculate** to start the calculation.

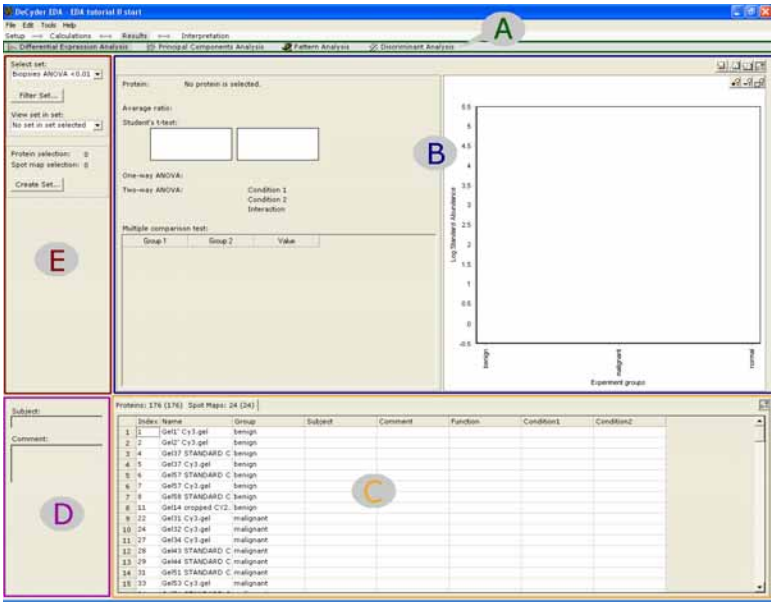
When the calculation has finished, this is indicated by the icon in front of the calculation and in the **Calculation status** field.

16.9.2 View and analyze the results of the PCA calculation

1 Select the **Results** step in the workflow area.



The **Results** window is displayed.



Select set: Biopsies ANOVA <0.01

View set in set: No set in set selected

Protein selection: 0
Spot map selection: 0

Protein: No protein is selected.

Average ratio:

Student's t-test:

One-way ANOVA:

Two-way ANOVA: Condition 1 Condition 2 Interaction

Multiple comparison test:

Group 1	Group 2	Value

Subject:

Comment:

Proteins: 176 (176) Spot Maps: 24 (24)

	Index	Name	Group	Subject	Comment	Function	Condition1	Condition2
1	1	Gel1' Cy3.gel	benign					
2	2	Gel2' Cy3.gel	benign					
3	4	Gel37 STANDARD C	benign					
4	5	Gel37 Cy3.gel	benign					
5	6	Gel57 STANDARD C	benign					
6	7	Gel57 Cy3.gel	benign					
7	8	Gel58 STANDARD C	benign					
8	11	Gel14 cropped CY2.	benign					
9	22	Gel31 Cy3.gel	malignant					
10	24	Gel32 Cy3.gel	malignant					
11	27	Gel34 Cy3.gel	malignant					
12	28	Gel43 STANDARD C	malignant					
13	29	Gel44 STANDARD C	malignant					
14	31	Gel51 STANDARD C	malignant					
15	33	Gel53 Cy3.gel	malignant					

The results window is divided into five main areas.

- **Results bar (A)**
  In this area, select the analysis results for display in the results view (B) and protein/spot map table (C) by clicking on the appropriate analysis.

- **Results view (B)**
  Shows detailed results for the selected protein/spot map in the protein/ spot map table for the current calculation.

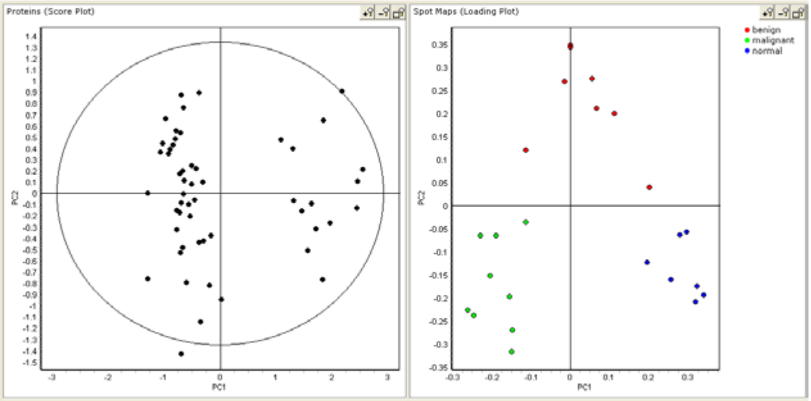
- **Protein/spot map table (C)** and **Protein/spot maps details area (D)**
  The protein/spot map table shows the result of all proteins and spot maps in a table format. The protein/spot map details area shows details of the selected protein/spot map in the results view or protein/spot map table.

- **Set area (E)**
  In this area new data sets can be created by selecting data directly in the results view and/or protein/spot map table or by filtering the data.

2 Click **Principal Component Analysis** in the Results bar (if not already displayed).

The result from the calculation is displayed in the Results view. The **Calculation result** field shows the name of the calculation for which results are displayed.

Note: *For detailed information about PCA, see Appendix D.*



3 **Look at the Score plot**
The score plot shows an overview of the proteins. The ellipse represents a 95% significance level. Proteins outside of the ellipse are outliers and should be checked. Outliers can be strongly differentially expressed proteins or mismatches in BVA. In this case, only one outlier is present and this protein has been checked in BVA when designing the tutorial. It is a strongly differentially expressed protein. Proceed with step 4.

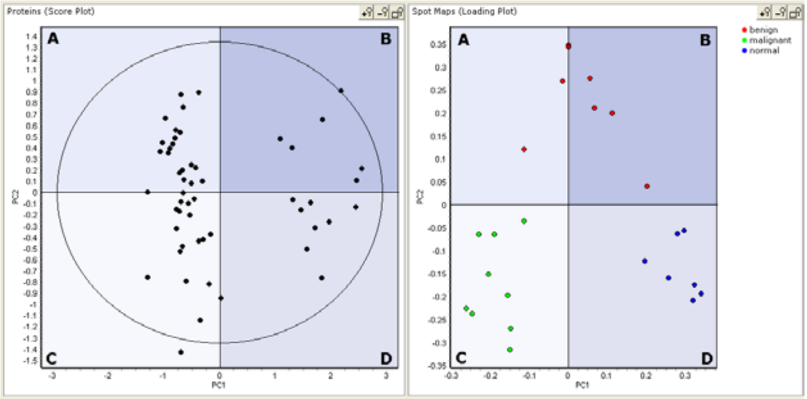
4 **Look at the Loading plot**
The loading plot shows the spot maps. The colors for experimental groups are displayed in the plot and a color legend with group names is displayed at the top-right corner.

5 **Compare the two plots**

It is possible to make a **rough** comparison of the two plots and **estimate** which proteins are up- or down-regulated in the different spot maps. Proteins and spot maps located in the corresponding quadrants have a correlation.

A few examples:

- Proteins in quadrant A and B are probably up-regulated in the spot maps in quadrant A and B (benign group) and down-regulated in the spot maps in quadrant C and D (normal and malignant groups). Moreover, the proteins in quadrant B are probably more up-regulated in the benign spot maps in quadrant B than in quadrant A.
- Proteins in quadrant C are probably up-regulated in spot maps in quadrant C (normal group) and down-regulated in spot maps in quadrant D (malignant group) and vice versa.



6 The analysis of the PCA calculation of the **Biopsies ANOVA < 0.01** set has been completed. The data looks OK and the protein outlier that was found is OK. Therefore, no new set needs to be created where outliers have been removed.

16.10Perform discriminant analysis

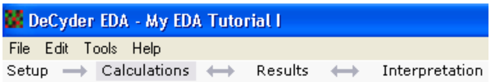
An overview of the data has been produced giving a view of the groupings and data relations. No protein mismatches were found and therefore no new set was created.

Discriminant analysis is now going to be performed. The analysis is divided into three sub-analyses:

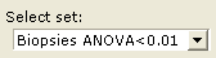
- **Marker selection**
  This analysis is used to find a set of proteins (biomarkers) that can be used to discriminate between experimental groups, i.e. the normal, benign and malignant biopsies. It is performed on the **Normal**, **Benign** and **Malignant** groups.
- **Classifier creation**
  When a set of biomarkers has been found, these will be used to create a classifier that will be used to classify the spot maps in the **Unknown** group.
- **Classification**
  Once the classifier has been created, classify the spot maps in the **Unknown** group.

16.10.1 Set up the Marker Selection calculation

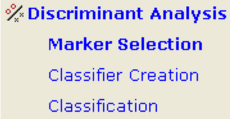
1 Click **Calculations** in the workflow area.



The **Calculations window** is displayed(see Fig. 16-2 for screenshot).

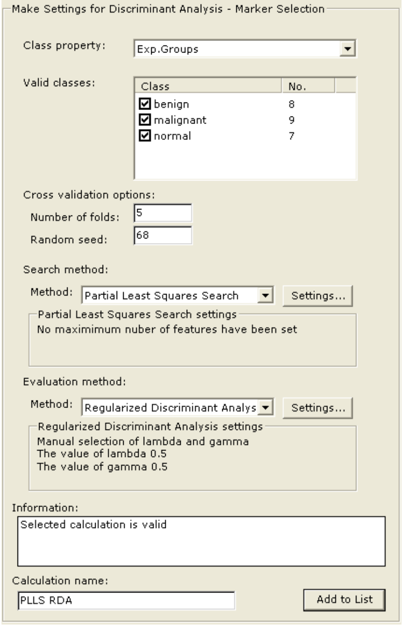

Select set: Biopsies ANOVA<0.01

2 Make sure that the **Biopsies ANOVA < 0.01** set is selected in the **Select set** field.



3 Select **Marker Selection** in the **Select calculation** area. The settings for the calculation is displayed to the right.

4 Enter the following settings for the first calculation:


Make Settings for Discriminant Analysis - Marker Selection

Class property: Exp.Groups

Valid classes:

Class	No.
☑ benign	8
☑ malignant	9
☑ normal	7

Cross validation options:
Number of folds: 5
Random seed: 68

Search method:
Method: Partial Least Squares Search
Partial Least Squares Search settings
No maximimum nuber of features have been set

Evaluation method:
Method: Regularized Discriminant Analys
Regularized Discriminant Analysis settings
Manual selection of lambda and gamma
The value of lambda 0.5
The value of gamma 0.5

Information:
Selected calculation is valid

Calculation name: PLLS RDA

a. Select **Exp. Groups** in the **Class property** area.

b. The **Normal**, **Benign** and **Malignant** boxes are checked by default in the **Valid classes** area. Use these settings.

c. In the **Cross validation options** area, make sure that **5** is entered in the **Number of folds** field and enter **68** in the **Seed** field.

d. In the **Search method** area, select **Partial Least Squares Search in** the **Method** drop-down list. Use the default settings displayed in the **Partial Least Squares Search settings** area (as displayed in the screenshot).

e. In the **Evaluation method** area, select **Regularized Discriminant Analysis** and use the default settings (as displayed in the screenshot).

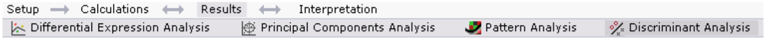
f. Type in *"PLSS RDA"* in the **Calculation name** field.

5 Click **Add to List** to add the calculation to the **Calculation List** to the right.

6 Click **Calculate** to start the calculation. When the calculation has finished, proceed with the next section .

16.10.2 View the results and create new sets with the biomarkers

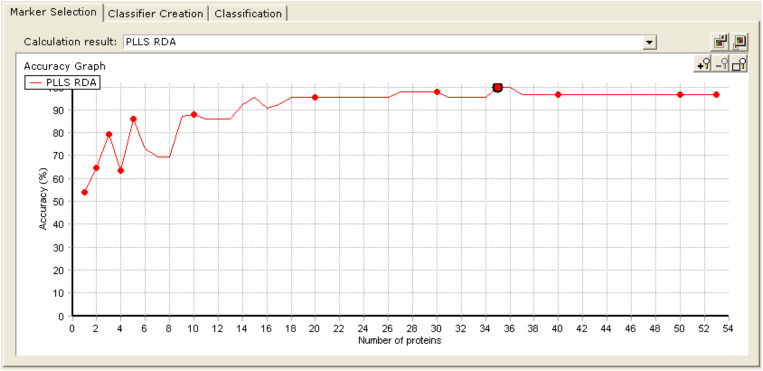
1 Select the **Results** step in the workflow area and then **Discriminant Analysis** in the Results bar.



2 Select the **Marker Selection** tab.

3 Select **PLLS RDA** from the **Calculation result** drop-down list.

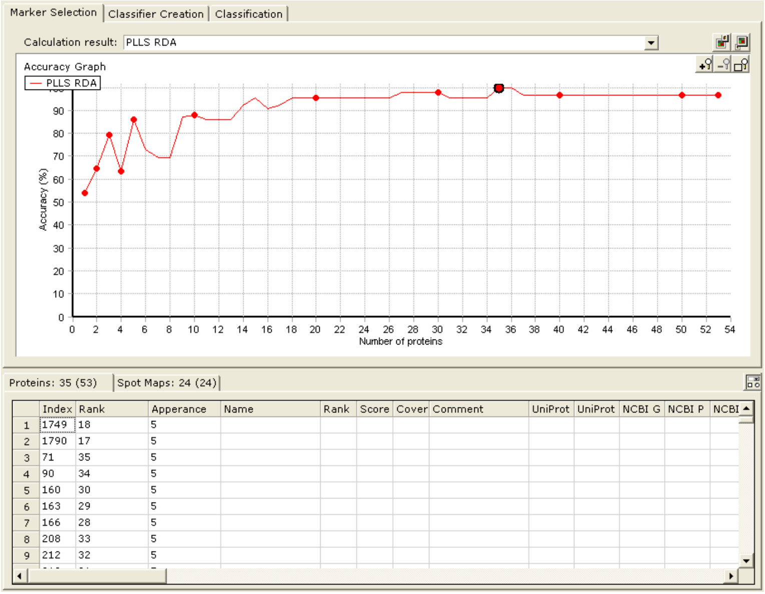
4 The **Results** window is displayed, showing the results of the feature selection calculation in the Results view.



Calculation result: PLLS RDA

The accuracy graph shows the accuracy of class prediction for different number of proteins. In this case the number of proteins that can discriminate between the different classes consists of 35 proteins (gives 100% accuracy).

5 Select the lowest number of proteins that give a 100% accuracy score for discriminating between the three groups (in this case 35) by clicking on it in the accuracy graph.



Calculation result: PLLS RDA

Proteins: 35 (53) | Spot Maps: 24 (24)

	Index	Rank	Apperance	Name	Rank	Score	Cover	Comment	UniProt	UniProt	NCBI G	NCBI P	NCBI
1	1749	18	5										
2	1790	17	5										
3	71	35	5										
4	90	34	5										
5	160	30	5										
6	163	29	5										
7	166	28	5										
8	208	33	5										
9	212	32	5										

35 proteins are shown in the Protein Table. As 5 folds were used in the Marker selection settings, five classifiers were created. Two parameters determine the quality of the result:

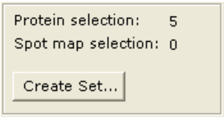
- **Appearance**
  For each protein, the number of classifiers that have selected this protein are listed in the **Appearance** column.

  If 5 folds were used, 5 in the **Appearance** column means that all classifiers have selected this protein and 1 means that only one classifier has selected the protein. It is primarily this parameter that is used to determine the quality of the results.

  In this case 5 folds were used, and 5 in the **Appearance** column means that all classifiers have selected the same protein (very good quality of the results).

- **Rank**
  For each protein, a rank value is displayed in the **Rank** column. This value is the mean of the different classifier ranking of the protein. Each classifier gives the first protein that is selected by the search method and gives the best result in the evaluation method rank 1, the second protein that is selected rank 2 and so on.

6 We will use all proteins in the Protein table to create the set with biomarkers, so select all proteins and click **Create set...** in the **Set** area. The **Create set** dialog is displayed.

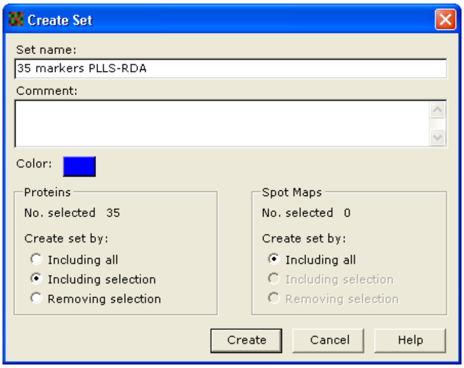


Protein selection: 5
Spot map selection: 0

7 Enter "*35 markers PLLS-RDA*" in the **Set name** field.

8 Select the **Including selection** radio button in the **Proteins** area and the **Including all** radio button in the **Spot Maps** area.

9 Click **Create**. The **35 markers PLLS-RDA** set is created and will be displayed in the **Select set** field.



Set name: 35 markers PLLS-RDA

Comment:

Color: 🟦

Proteins
No. selected 35
Create set by:
○ Including all
● Including selection
○ Removing selection

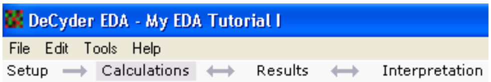
Spot Maps
No. selected 0
Create set by:
● Including all
○ Including selection
○ Removing selection

16.10.3 Set up the classifier creation calculations

A new set, 35 markers PLLS-RDA, has been created. Now, a classifier will be built using these markers, that will be able to classify new samples into the correct groups.

Set up the calculation as follows:

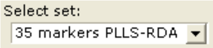
1 Click **Calculations** in the workflow area.



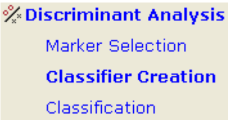
DeCyder EDA - My EDA Tutorial I
File Edit Tools Help
Setup ⟶ Calculations ⟷ Results ⟷ Interpretation

The **Calculations window** is displayed(see Fig. 16-2 for screenshot).

2 Make sure that the **35 markers PLLS-RDA** set is selected in the **Select set** field.



Select set: 35 markers PLLS-RDA



3 Select **Classifier creation** in the **Select calculation** area. The settings for the calculation are displayed to the right.

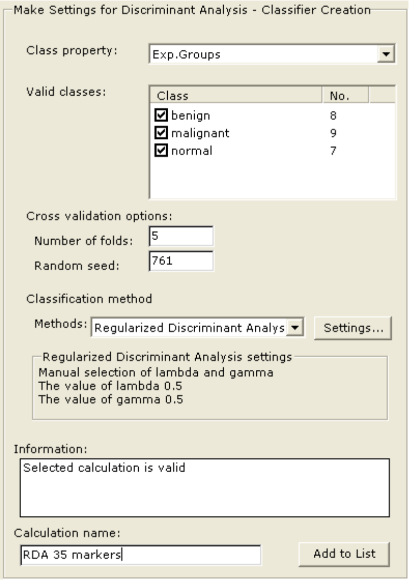
4 Enter the following settings:

a. Select **Exp. Groups** in the **Class property** area.

b. Make sure that the **Normal**, **Benign** and **Malignant** boxes in the **Valid classes** area are checked.

c. In the **Cross validation options** area, enter **5** in the Number of folds field and enter **761** in the **Seed** field.

d. Select **Regularized Discriminant Analysis** in the **Classification method** area and use the default settings (as displayed in the screenshot).



Make Settings for Discriminant Analysis - Classifier Creation

Class property: Exp.Groups

Valid classes:

Class	No.
☑ benign	8
☑ malignant	9
☑ normal	7

Cross validation options:

Number of folds: 5

Random seed: 761

Classification method

Methods: Regularized Discriminant Analys

Regularized Discriminant Analysis settings
Manual selection of lambda and gamma
The value of lambda 0.5
The value of gamma 0.5

Information:
Selected calculation is valid

Calculation name: RDA 35 markers

e. Type in *"RDA 35 markers"* in the **Calculation name** field.

5 Click **Add to List** to add the calculation to the **Calculation List**.

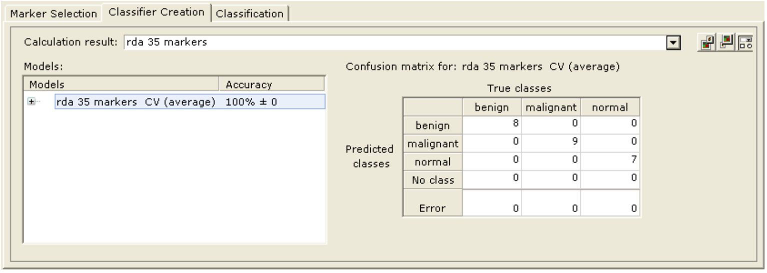
6 Click **Calculate** to start the calculation.

During calculation the status of each calculation is indicated by an icon in front of the calculation and the progress of the calculation is displayed by a progress bar. The status of the currently selected calculation is also displayed in the **Status of selected list item** area.

16.10.4 View the results

1 Select the **Results** step in the workflow area, select **Discriminant Analysis** in the Results bar and select the **Classifier Creation** tab in the Results view.

2 Select **RDA 35 markers** from the **Calculation result** drop-down list. The results are displayed in the Results view.



Calculation result: rda 35 markers

Models:

Models	Accuracy
rda 35 markers CV (average)	100% ± 0

Confusion matrix for: rda 35 markers CV (average)

		True classes		
		benign	malignant	normal
Predicted classes	benign	8	0	0
	malignant	0	9	0
	normal	0	0	7
	No class	0	0	0
	Error	0	0	0

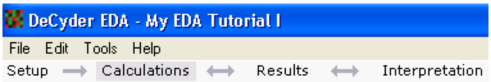
3 In the **Models** area, select the created classifier **RDA 35 markers CV (average)**. The accuracy of class prediction is 100% ± 0. The highest accuracy with smallest variation possible is desirable.

4 In the **Confusion matrix** area to the right an overview of the classification of the spot maps is displayed. Spot maps that were wrongly classified are displayed in red.

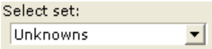
16.10.5 Set up the classification calculation

A classifier has been created. Classification of the samples in the **Unknown** group using the created classifier will now be performed.

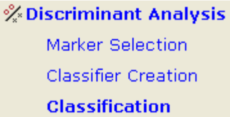
1 Click **Calculations** in the workflow area.



The **Calculations window** is displayed(see Fig. 16-2 for screenshot).

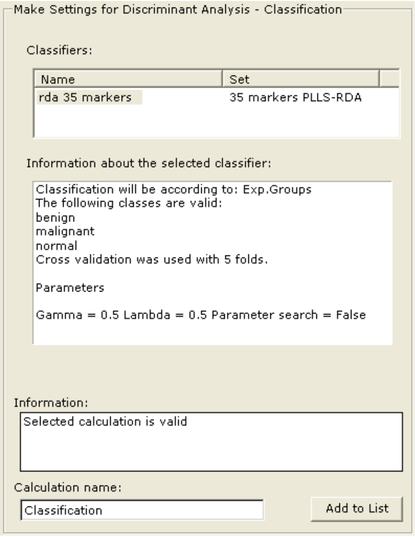

Select set:
Unknowns

2 Make sure that the **Unknowns** set is selected in the **Select set** field.



3 Select **Classification** in the **Select calculation** area. The settings for the calculation are displayed to the right.

4 Enter the following settings:


Make Settings for Discriminant Analysis - Classification

Classifiers:

Name	Set
rda 35 markers	35 markers PLLS-RDA

Information about the selected classifier:

Classification will be according to: Exp.Groups
The following classes are valid:
benign
malignant
normal
Cross validation was used with 5 folds.

Parameters

Gamma = 0.5 Lambda = 0.5 Parameter search = False

Information:
Selected calculation is valid

Calculation name:
Classification

a. Select the **rda 35 markers** classifier in the **Classifiers** area.

Information on the classifier is displayed in the **Information about the selected classifier** area.

b. Type in *"Classification"* in the **Calculation name** field.

5 Click **Add to List** to add the calculation to the **Calculation List** to the right.

6 Click **Calculate** to start the calculation.

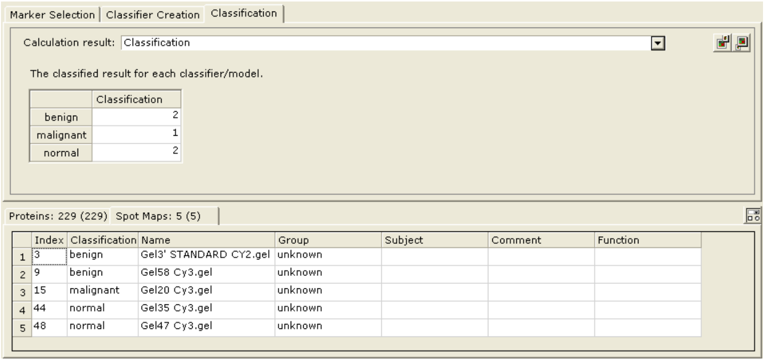
7 When the calculation has finished, proceed with the next section.

16.10.6 View the results of the classification

The result of the classification will be compared to the results of the classifications by pathologists.

1 Select the **Results** step in the workflow area, select **Discriminant Analysis** in the Results bar and select the **Classification** tab in the Results view.

2 Select **Classification** from the **Calculation result** drop-down list. The results are displayed in the Results view and the Spot Map Table.



Calculation result: Classification

The classified result for each classifier/model.

	Classification
benign	2
malignant	1
normal	2

Proteins: 229 (229) Spot Maps: 5 (5)

	Index	Classification	Name	Group	Subject	Comment	Function
1	3	benign	Gel3' STANDARD CY2.gel	unknown			
2	9	benign	Gel58 Cy3.gel	unknown			
3	15	malignant	Gel20 Cy3.gel	unknown			
4	44	normal	Gel35 Cy3.gel	unknown			
5	48	normal	Gel47 Cy3.gel	unknown			

The number of spot maps classified into the different groups are displayed in the Results view. The Spot Map Table shows the results of the classification for each spot map in the Classification column.

3 It is now possible to compare the classification results with the results of the classification by pathologists. The results are shown in the table below and show that the spot maps have been classified correctly.

Spot map	Class prediction by EDA	Class prediction by pathologists
Gel3' STANDARD CY2.gel	Benign	Benign
Gel58 Cy3.gel	Benign	Benign
Gel20 Cy3.gel	Malignant	Malignant
Gel35 Cy3.gel	Normal	Normal
Gel47 Cy3.gel	Normal	Normal

Appendix A Normalization

A.1 Overview

Normalization in EDA can be performed in three different ways:

- *Workspace normalization*
  Normalization between the imported BVA workspaces will correct for non-biological variation between the workspaces. It is only necessary to perform if several BVA workspaces that do not use the same internal standard (but have used the same Master or a Template for matching) are included in the EDA workspace.

  Note: *All spots do not need to be matched, as is the case if linking with Template, to perform normalization.*

- *Scaling*
  Scaling can be used to rescale the data to an experimental group instead of to the internal standard. The log standard abundance values for proteins on spot maps in other groups are then compared to the log standard abundance values for proteins in the experimental group (which is zero) instead of to the standard.

- *Standardization*
  Standardization normalizes the data so that all proteins and/or spot maps have a mean of 0 and standard deviation of 1.

Several normalization methods can be performed after each other.

A.2 Workspace Normalization

If two or more BVA workspaces in the EDA workspace use different internal standards (but have used the same Master or a Template for matching), normalization between the BVA workspaces should be performed.

Normalization between the BVA workspaces is performed to remove system variation so that the protein expression values (i.e. the biological variation) can be compared in the different BVA workspaces.

A.2.1 Using a common experimental group for normalization

If a common experimental group, for example a control group, exists in the imported BVA workspaces, it can be used for normalization between the different BVA workspaces. A linear regression model is then created between each BVA workspace and a reference BVA workspace.

The procedure fits a model for each BVA workspace by using the log standard abundance of the spots in the common experimental groups that are also matched between the reference BVA workspace and the BVA workspace.

The result of each model is a slope which is the normalizing factor between the reference BVA workspace and the BVA workspace to normalize. All data points in the BVA workspace to be normalized are then multiplied with this normalizing factor.

Since a new model is created for each BVA workspace, the different BVA workspaces are normalized with different normalizing factors.

A.2.2 Using a reference BVA workspace for normalization

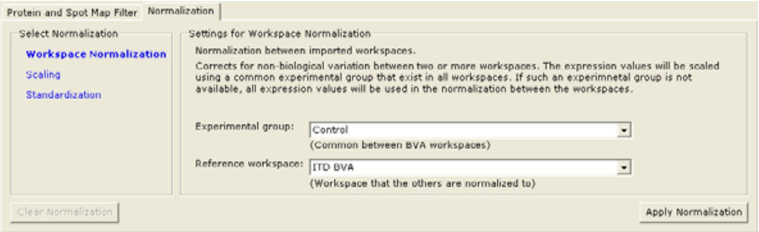
If a common experimental group does not exist among the workspaces, normalization can still be performed.

The difference in using a common experimental group is that all matched spots between the reference BVA workspace and the BVA workspace to normalize in all spot maps are used when fitting a model for the BVA workspace.

Note: *It is recommended to have a common experimental group if several BVA workspaces with different internal standards are to be linked.*

A.2.3 Performing workspace normalization in EDA

1 Select **Workspace Normalization** in the **Manual Base Set Creation** dialog.



Settings for Workspace Normalization

Normalization between imported workspaces.

Corrects for non-biological variation between two or more workspaces. The expression values will be scaled using a common experimental group that exist in all workspaces. If such an experimnetal group is not available, all expression values will be used in the normalization between the workspaces.

Experimental group: Control
(Common between BVA workspaces)

Reference workspace: ITD BVA
(Workspace that the others are normalized to)

2 *If a common experimental group exists among the BVA workspaces*, select the appropriate one to use in the normalization from the **Experimental group** drop-down list.

3 Select a **Reference workspace** in the drop-down list with which the other workspaces should be normalized.

4 Click **Apply Normalization** to normalize the data. The heat map will be updated with the new values.

To clear the normalization, click **Clear Normalization.**

A.3 Scaling

Scaling can be used to rescale the data to an experimental group (for example a Control group) instead of to the internal standard. The log standard abundance values for proteins on spot maps in other groups are then compared to the log standard abundance values for proteins in the common experimental group (which is zero) instead of to the standard.

After the normalization, a zero log standard abundance indicates that the protein is not up- or down-regulated compared to the common experimental group.

When performing scaling, the log standard abundance mean for each protein is calculated for the spot maps in the common group in each workspace. The mean for each protein is then subtracted from the corresponding spots in all groups in the workspace.

$$y_{ij} = x_{ij} - \bar{x}_{ic}$$

$$\bar{x}_{ic} = \frac{1}{n}\sum_{c=1}^{n} x_{ic}$$

where y_{ij} is the normalized log standard abundance for protein **i** on spot map **j**

x_{ij} is the log standard abundance for protein **i** on spot map **j**

$\bar{x}_{ic}$ is the mean log standard abundance for protein **i** on the spot map **c** among the common experimental group spot maps

x_{ic} is the log standard abundance for protein **i** on the spot map **c** among the common experimental group spot maps

Fig. A-1 shows the principle for normalization using a common experimental group (in this case the Control group). In the figure, the mean of all proteins on the spot maps is displayed on the y-axis.

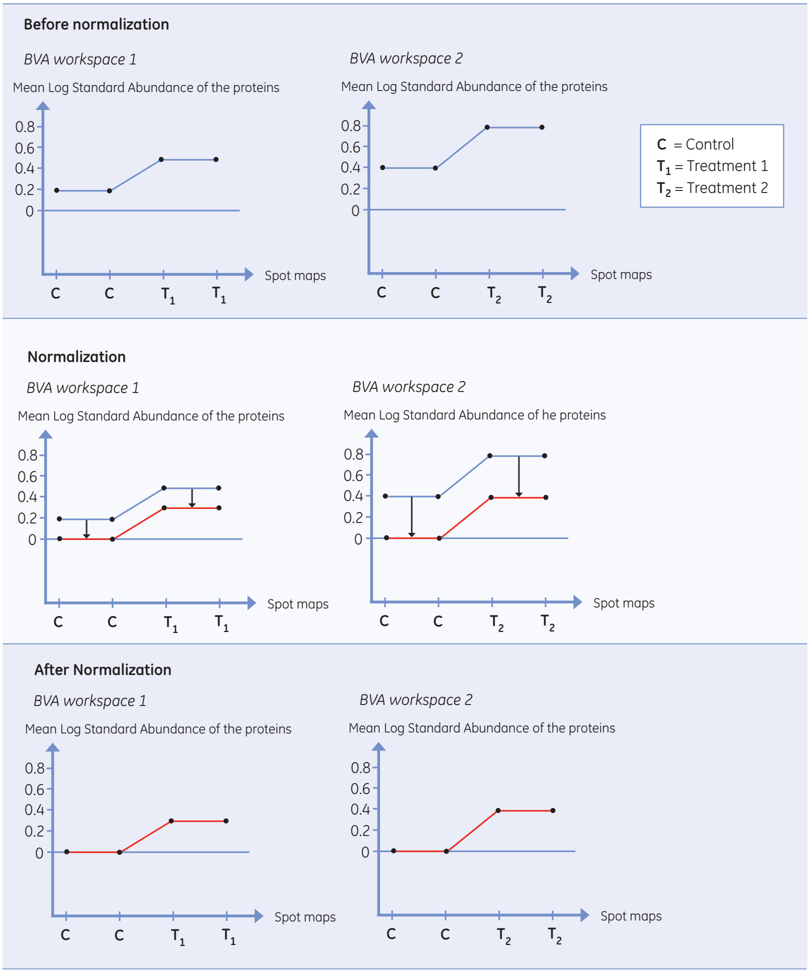
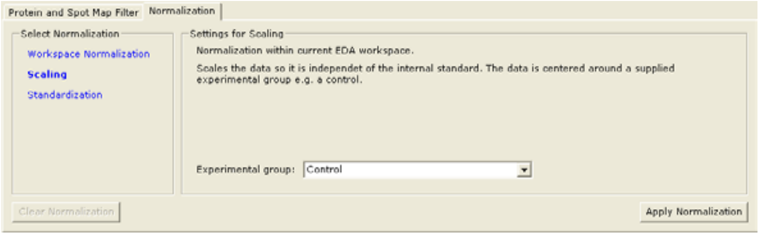

Before normalization
BVA workspace 1
Mean Log Standard Abundance of the proteins
0.8
0.6
0.4
0.2
0
Spot maps
C
C
T1
T1
BVA workspace 2
Mean Log Standard Abundance of the proteins
Spot maps
C
C
T2
T2
C = Control
T1 = Treatment 1
T2 = Treatment 2
Normalization
BVA workspace 1
Mean Log Standard Abundance of the proteins
BVA workspace 2
Mean Log Standard Abundance of he proteins
After Normalization
BVA workspace 1
Mean Log Standard Abundance of the proteins
BVA workspace 2
Mean Log Standard Abundance of the proteins

Fig A-1. The principle for normalization using a common experimental group.

A.3.1 Performing Scaling in EDA

1 Select **Scaling** in the **Manual Base Set Creation** dialog.



Normalization within current EDA workspace.

Scales the data so it is independet of the internal standard. The data is centered around a supplied experimental group e.g. a control.

Experimental group: Control

2 Select the group, around which to center the data, from the **Experimental group** drop-down list.

3 Click **Apply Normalization** to normalize the data. The heat map will be updated with the new values.

A.4 Standardization

This method standardizes the data so that all proteins and/or spot maps have the mean of 0 and standard deviation of 1. This can be useful for advanced data mining applications.

Four methods can be used for standardization of the data:

- Mean centering on protein
- Mean centering on spot map
- Standard deviation on protein
- Standard deviation on spot map

One, several, or all methods can be used, in any order. However, different method orders will give different results. The reason is that each calculation is based on the results of the previous calculation.

A.4.1 Mean centering on protein

This method standardizes each protein (represented as a row in the expression matrix) by calculating the mean of each protein's expression in all spot maps and then subtracting each value for the protein by the mean. The mean of a protein (row) will be zero, N(0,s).

$$y_{ij} = x_{ij} - \bar{x}_{ic}$$

$$\bar{x}_{ic} = \frac{1}{n}\sum_{c=1}^{n} x_{ic}$$

where y_{ij} is the standardized log standard abundance for protein **i** on spot map **j**

x_{ij} is the log standard abundance for protein **i** on spot map **j**

$\bar{x}_i$ is the mean log standard abundance for protein **i**

A.4.2 Mean centering on spot map

This method standardizes each spot map (represented as a column in the expression matrix) by calculating the mean of expression of all of the proteins on the spot map and then subtracting the value for the spot map by the mean. The mean of a spot map (column) will be zero, N(0,s).

A.4.3 Standard deviation on protein

This method standardizes each protein (represented as a row in the expression matrix) by calculating the standard deviation of each protein's expression on the selected spot maps and then dividing each value for the protein with the standard deviation. The standard deviation of a protein (row) will be one, N(m,1).

$$y_{ij} = \frac{x_{ij}}{\sigma_i}$$

$$\sigma_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2}$$

where y_{ij} is the standardized log standard abundance for protein **i** on spot map **j**

x_{ij} is the log standard abundance for protein **i** on spot map **j**

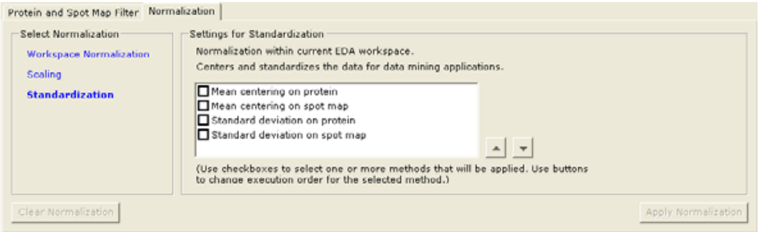
σ_i is the standard deviation for protein **i**

A.4.4 Standard deviation on spot map

This method standardizes each spot map (represented as a column in the expression matrix) by calculating the standard deviation for expression of all of the proteins on the spot map and then dividing each value for the spot map with the standard deviation. The standard deviation of a spot map (column) will be one, N(m,1)

A.4.5 Performing standardization in EDA

1 Select **Standardization** in the **Manual Base Set Creation** dialog.



Normalization within current EDA workspace.
Centers and standardizes the data for data mining applications.

□ Mean centering on protein
□ Mean centering on spot map
□ Standard deviation on protein
□ Standard deviation on spot map

(Use checkboxes to select one or more methods that will be applied. Use buttons to change execution order for the selected method.)

2 Check the appropriate boxes to select the methods to use in the standardization.

3 If required, change the order in which to apply the methods on the data by selecting a method and clicking the appropriate arrow to move the method up or down in the list.

> ***Note:*** *Different orders of the calculations will give different results. The reason is that each calculation is based on the results of the previous calculation.*

4 Click **Apply Normalization** to normalize the data. The heat map will be updated with the new values.

Appendix B Statistics and algorithms - Introduction

B.1 Overview

B.1.1 Purpose

The DeCyder EDA module encompasses a broad range of statistical methods which are applied to biological data. Although most of the analyses can be performed without advanced statistical knowledge, this appendix gives the user a possibility to gain a deeper insight of the underlying methods and algorithms used. It also serves as a reference for advanced users in need of defined mathematical formulas and literature references.

B.1.2 Outline

Appendices C-F describe the analyses available in the EDA software. In each section, one analysis is described. The analyses are arranged to correspond with the calculations view in EDA.

Each analysis chapter contains a brief overview of the methods available for performing the analysis, the advantages and disadvantages of each method and a description of cases in which one method may be preferable to another.

Each analysis section also contains subsections describing each method in detail, putting the method settings into mathematical context. The method subsections also include literature references.

Appendix C Statistics and algorithms - Differential Expression Analysis

C.1 Overview

Differential expression analysis is used to identify proteins with significant differences in expression between two or more predefined groups. The groups could for example contain samples made at different time points or subjects given different diagnosis, or treated with different drug doses.

Method	Function	Application
Average Ratio	Calculates the average difference in expression between two groups. Also known as fold change.	Is used for determining protein fold change between two different groups.
Student's T-test	Calculates the significance in expression difference between two groups.	Is used for identifying the significantly differentially expressed proteins between two groups.
One-way ANOVA	Calculates the significance in expression difference between several groups.	Is used for determining the significantly expressed proteins between several groups. One-way ANOVA is often used as a filter for decreasing the number of proteins before any other statistical calculations in EDA.
Two-Way ANOVA	Calculates the significance in expression difference for two conditions, *e.g.* time and dose and their interaction.	Is used for determining the significantly expressed proteins over two conditions and the interaction between the conditions

Table C-1. Overview of methods used in differential expression analysis.

Method	Function	Application
False Discovery Rate	Adjusts the results of the differential expression analysis to account for occurrences of false positives.	Is used for adjusting the p-values of significance tests due to false positives.
Multiple Comparison Test (One-way ANOVA)	Calculates a pair-wise measure between all experimental groups, showing if the difference in protein expression is significant.	Is used for identifying between which groups a protein is significantly differentially expressed when running One-way ANOVA.

Table C-2. Overview of supplementary methods used in differential expression analysis.

C.2 Concepts

C.2.1 Independent and Paired tests

The different samples in an experiment can be either independent or paired.

Independent samples originate from separate samples that contain different sets of individual subjects and are most commonly used. Paired samples are present when each sample in one group corresponds to a matching sample in the other group(s). A typical example would be the same group of patients before and after a treatment. All of the differential expression analysis tests in EDA are implemented both as independent and paired tests.

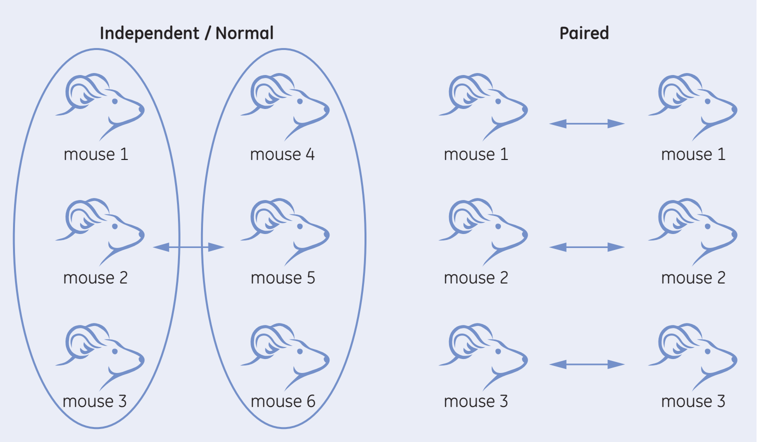


Fig C-1. Design of independent and paired tests.

C.2.2 Null Hypothesis

When performing a statistical hypothesis test one always tests against a null hypothesis, which essentially is the opposite of what one expects.

For example if one is interested in if the means of the two groups are different, the null hypothesis is that they are the same.

C.2.3 Sampling Distribution

In order to uncover if a certain hypothesis test result is significant one has to test against the corresponding sampling distribution. A sampling distribution is a model of a distribution of sample statistics.

For example, suppose that a sample size of ten (N=10) is taken from some population. The mean of the ten numbers is computed. Then a new sample of ten is taken and the mean is again computed. If this were to be repeated an infinite number of times, the distribution of the now infinite number of sample means would be called the sampling distribution of the mean.

This results in every statistic having a sampling distribution.

For example, the sampling distribution of the differences in two means (if samples are large) has the t-distribution. Therefore, the t-distribution is used to evaluate if the observed difference in means is statistically significant.

So why is the t-distribution used instead of the normal distribution?

The difference between two means is normally distributed for large samples and the t-distribution approximates this normal distribution in large samples. For small samples, the distribution of differences in the mean is not quite normal and the normal distribution cannot be used. A new distribution is needed. This was noted by a quality control statistician at Guinness Brewing (W.S. Gossett), but because the brewery didn’t allow the employees to publish their work Gossett published under the name “Student” (see Student’s T-test below).

C.2.4 P-Value

The probability value (p-value) of a statistical hypothesis test is the probability of getting a result as extreme or more extreme than the one observed if the proposed null hypothesis is correct.

A small p-value provides evidence against the null hypothesis, because data have been observed that would be unlikely if the null hypothesis were correct. Thus, the null hypothesis can be rejected when the p-value is sufficiently small.

The p-value is often compared to a significance level which is a fixed probability of wrongly rejecting the null hypothesis, if it is true. The significance level is also the probability of rejecting the null hypothesis wrongly (type I error). By chance a % of all test cases falsely reject the null hypothesis.

Therefore, the significance level should be as small as possible in order to protect the null hypothesis and to prevent, as far as possible, the investigator from inadvertently making false claims.

Example

If a statistical test returns a p-value < a, for example 0.001 if the significance level a is set to 0.05 the null hypothesis should be rejected.

In an experiment where 1000 p-values are generated and the same significance level is applied, 1000* 0.05 = 50 test cases have falsely rejected the null hypothesis due to stochastic changes.

Usually, the significance level is chosen to be 0.05, 0.01 or even 0.001.

C.3 Average Ratio

C.3.1 Introduction

Average Ratio is calculated to investigate the protein expression fold change between two groups. The average ratio value indicates the standardized volume ratio between the two groups or populations.

Example

In a control/treated experiment it might be interesting to calculate the fold change between the two groups for all proteins. Examples of expression values for a protein are shown below. The average ratio can be used to calculate protein abundance in the treated samples.

Protein	Control sample	Control sample	Control sample	Treated sample	Treated sample	Treated-sample
1	0.24	0.25	0.32	-0.22	-0.23	-0.28

Table C-3. The average ratio (Treated/Control) = -3.26614, which means that the protein abundance is 3.26 times less in the treated samples than in the control samples.

C.3.2 Detailed description

Independent Samples

The average ratio for independent samples are calculated as follows:

$$\text{Average Ratio} = m_a/m_b$$

where **$m_{a,b}$** is the mean of the expression values in group **a** or **b**. Note that the expression values are not logged.

Paired Samples

The average ratio for paired samples is calculated as follows:

$$\text{Average Ratio} = \sum_{s} (m_{as} - m_{bs})$$

where **$m_{as,bs}$** is the mean of expression values in group **a** or **b** for subject (individual) **s**.

Values are displayed as fold change, so decreases in expression are in the range of -∞ to -1 and increase in the region +1 to +∞. Hence a two-fold increase or decrease is represented by 2 or -2, respectively (not 2 and 0.5).

C.4 Student's T-test

C.4.1 Introduction

Student's T-test, often known simply as the T-test, is one of the most commonly used of all statistical tests. The Student's T-test is used to test whether a variable differs between two groups. The Student's T-test in EDA is performed as an equal variance two-tailed test, since the direction of change (i.e. increase and decrease) in the standardized abundance parameter is considered.

Example

If the same example as in the Average ratio case above is used, the p-value for the protein is: 0.0000806, thus the null hypothesis can be rejected at a very low significance level.

The protein is significantly differentially expressed between the treated and the control group.

C.4.2 Detailed description

Independent Samples

The Student's T-test for independent samples is calculated as follows

$$t = \frac{\mu_a - \mu_b}{\sigma_{a-b}}$$

where μ_a-μ_b is the difference in means between two groups.

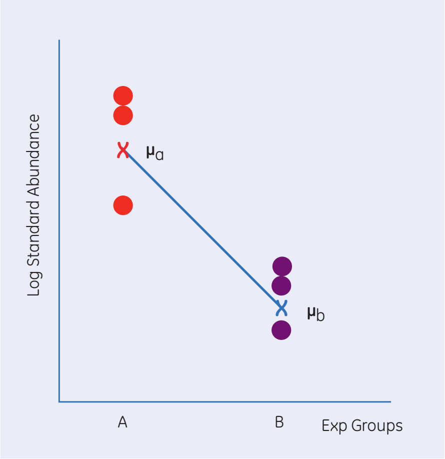


Fig C-2. Illustration of means in groups used in T-test calculation.

Since equal variance is assumed, the deviation term is calculated as

$$\sigma_{a-b} = \sqrt{\left(\frac{SS_a + SS_a}{N_a + N_b - 2}\right)\left(\frac{1}{N_a} + \frac{1}{N_b}\right)}$$

where **N_x** is the number of values in group **x** and **SS_x** is the sum of squares in group **x**:

$$SS = \sum (x_i - \mu_x)^2$$

The t-value is then compared to the t distribution with a degree of freedom **df** equal to:

$df = N_a + N_b - 2$

Paired Samples

The Student's T-test for paired samples (repeated measures) is calculated as follows:

$$t = \frac{\mu_D}{\sigma_D}$$

where μ_D is the mean of the difference for the individual in both groups.

$$\mu_D = \frac{1}{N}\sum_i D_i$$

where **N** is the number of subjects (individuals).

$$D_i = x_{ai} - x_{bi}$$

where $\mathbf{x_{ai}}$ and $\mathbf{x_{bi}}$ are the values for subject **i** in group **a** or **b**, as shown in Fig. C-3.

$$SS = \sum (x_i - \mu_x)^2$$

where $\mathbf{SS_D}$ is the sum of squares of the $\mathbf{D_i}$ values

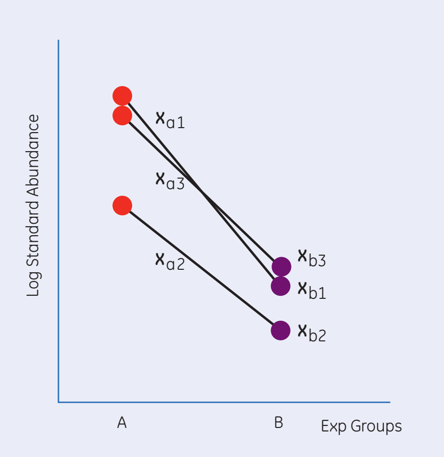


Fig C-3. Conceptual example of protein expression values in experimental groups.

The t-value is then compared to the t distribution with a degree of freedom **df** equal to:

$df = N - 1$

C.5 One-Way ANOVA

C.5.1 Introduction

Analysis of Variance (ANOVA) is one of the most important statistical tests available for biologists and is essentially an extension of the logic of Student's T-tests to those situations where the comparison of the means of several groups is required. Thus, when comparing two means, ANOVA will give the same results as the T-test for independent samples (if comparing two different groups or observations). There are no restrictions on the number of groups that can be analyzed. It is equally valid for testing differences between two groups as among twenty.

Independent Samples

In an experiment including 4 treatments (A, B, C, D), 6 separate t-tests (comparing A with B, A with C, A with D, B with C, B with D and C with D) would be needed. If there were 10 treatments as much as 45 separate t-tests would be needed. This would be very time-consuming but would also be prone to errors since each t-test introduces a 1% chance of our conclusion being wrong (when testing for $p < 0.01$).

ANOVA overcomes this problem by detecting significant differences between the treatments as a whole. Therefore, there is only one test to test the differences between our treatments.

As with the Student's T-test the ANOVA tests can be either independent or paired. Paired ANOVA tests are also called repeated measures.

Repeated Measures

Repeated Measures (RM) One-Way ANOVA is a generalization of the paired Student's T-test in the same manner One-Way ANOVA is a generalization of Student's T-test.

A hypothesis for any number of treatments in the same subjects (individuals) can thus be tested, which is useful for time or dose series etc.

Example

In a study, expression levels of 3 different time points have been collected for 3 different subjects (individuals).

Subject	Time point 1	Time point 2	Time point 3
1	Sample 1	Sample 2	Sample 3
2	Sample 4	Sample 5	Sample 6
3	Sample 7	Sample 8	Sample 9

C.5.2 Detailed Description

This section describes the traditional way of performing One-Way ANOVA calculations. However, in DeCyder EDA, the ANOVA algorithms are implemented using multiple linear regression analysis to handle unbalanced data sets (groups with different sizes). It can be proved mathematically that analysis of variance and regression are two ways of calculating the same result.

For further information on the implementation, see the reference list.

Traditional Independent One-Way ANOVA

Since One-Way ANOVA can be said to be a generalization of Student's T-test the assumption is that there are two or more independent groups of measures, e.g. A, B, C. The question here is: Do the means of the groups significantly differ from one another?

Between Groups

If one compares to the Student's T-test there is a need for a measure of difference between the group means for the ratio nominator and since the difference between two means cannot be used (since there are three groups), a variance measure between the group means is used instead. The variance is measured by a sum of squares (SS) over each group and the total mean.

The sum of squares can be calculated as:

$$SS = \sum (x_i - \mu_x)^2$$

which gives

$$SS^* = \sum (\mu_i - \mu_{TOT})^2$$

where μ_{TOT} is the mean of all values. But **SS*** is not a sum of squared values since the means are not summed over all values, but all groups. However, if the value is multiplied with the group size, the complete sum of squares between the groups $\mathbf{SS_{bg}}$ is calculated.

$$SS_{bg} = \sum_i n_i (\mu_i - \mu_{TOT})^2$$

where $\mathbf{n_i}$ is the number of values in group **i**.

The measure of $\mathbf{SS_{bg}}$ is similar to the nominator of the T-test; $\mathbf{m_a - m_b}$ is the difference between two means; $\mathbf{SS_{bg}}$ is the difference among three or more means. A large $\mathbf{SS_{bg}}$ might lead to a significant p-value.

Within Groups

Continuing the comparison with Student's T-test, a corresponding value for the denominator σ_{a-b} needs to be found.

σ_{a-b} includes the random variability in each group and is calculated using the SS for the groups. When there are more than two groups SS can still be used. The residual variability within the groups can be written as:

$$SS_{wg} = \sum_i SS_i$$

where $\mathbf{SS_i}$ is the sum of squares value for group **i**.

The **SS**$_{wg}$ and **SS**$_{bg}$ account for variability observed within and between the treatment groups. In addition it is possible to calculate the total variability observed in the data by computing the sum for squares of all observations:

$$SS_{TOT} = \sum_i (x_i - \mu)^2$$

where µ is the grand mean of all the data and **x**$_i$ is the data point **i**.

The three sum of squares are related in a very simple way:

$SS_{TOT} = SS_{bg} + SS_{wg}$

Thus, the total variability, measured by the sum of squares can be portioned into the variability between groups and the variability within groups.

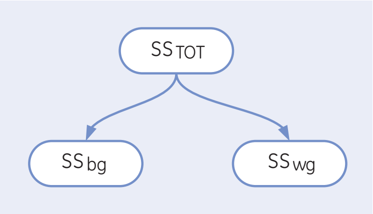


Fig C-4. Total variability can be portioned between and within groups.

Mean Square

Since the definition of sample variance is:

$$\sigma^2 = \frac{SS}{N-1} = \frac{SS}{df}$$

where **N** is the size of the sample and **df** is the degree of freedom.

In an ANOVA context, this variance is called a mean square and often denoted **MS**. Thus the between groups mean square is:

$$MS_{bg} = \frac{SS_{bg}}{df_{bg}}$$

where **df_{bg}** = N - 1, where **N** is the group size.

The within-groups **MS** are constructed similarly but each of the within-groups measures of **SS** are associated with a certain number of degrees of freedom, **N_x** - 1, respectively.

Therefore, the number of degrees of freedom associated with the composite within-groups measure **SS_{wg}** is:

$$df_{wg} = \sum_i (N_i - 1)$$

and the mean square:

$$MS_{wg} = \frac{SS_{wg}}{df_{wg}}$$

F-Ratio

Thus in analogy with the T-ratio in Student's T-test an F-ratio is calculated, in which the two MS values are combined as:

$$F = \frac{MS_{bg}}{MS_{wg}}$$

This value is then tested against the corresponding sampling distribution, the F distribution, with the two degrees of freedom **df_{bg}** and **df_{wg}**.

If the null hypothesis was that all groups were drawn from the same population, i.e. the treatments had no effect, then the two **MS** would be similar and the ratio would be close to 1.

In contrast, if the treatments had an effect then there is more variation between the means than within means and the ratio would be larger than 1.

Traditional Repeated Measures One-Way ANOVA

Some of the concepts of independent One-Way ANOVA are similar to those in RM One-Way ANOVA.

The total sum of squares $\mathbf{SS_{TOT}}$ can be divided into $\mathbf{SS_{bg}}$ and $\mathbf{SS_{wg}}$.

In RM ANOVA however part of the within-group variability depends on the individual variability and the $\mathbf{SS_{wg}}$ can thus be divided into a **SS** for the subject variability $\mathbf{SS_{subj}}$ and a **SS** for the random variability called $\mathbf{SS_{error}}$:

$$SS_{TOT} = SS_{wg} + SS_{bg}$$

where

$\mathbf{SS_{wg}} = SS_{subject} + SS_{error}$

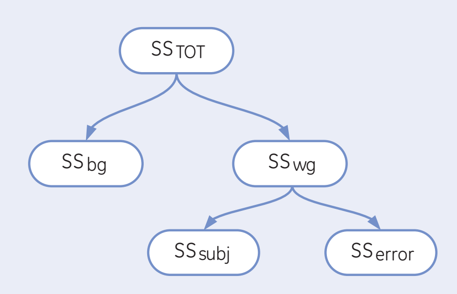


Fig C-5. Total variability can be portioned between and within group.

In analogy to the $\mathbf{SS_{bg}}$

$$SS_{bg} = \sum_i n_i(\mu_i - \mu_{TOT})^2 = \frac{\sum_i \left(\sum_j x_{ij}\right)^2}{N_i} - \frac{\left(\sum x\right)^2}{N_{tot}}$$

where $\boldsymbol{x_{ij}}$ is the value for group ***i*** and subject **j** and $\mathbf{N_i}$ is the size of the group **i**,

the $\mathbf{SS_{subj}}$ can be calculated as:

$$SS_{subj} = \frac{\sum_j \left(\sum_i x_{ij}\right)^2}{k} - \frac{\left(\sum x\right)^2}{N_{tot}}$$

where $\mathbf{x_{ij}}$ is the value for group ***i*** and subject **j** and **k** is the number of treatment groups.

So the $\mathbf{SS_{subj}}$ is the amount of variability within the total data that derives from individual differences among the subjects.

The within group variability would, if the variability from the subjects are left in the calculations, result in a wrong conclusion. The difference in group means for each subject needs to be removed before analyzing the results. For example, the difference in group means could depend on one subject alone and not the rest of the subjects, so the pre-existing differences between the subjects needs to be removed.

The difference is the $\mathbf{SS_{subj}}$.

Therefore, the F ratio is calculated as follows:

$$F = \frac{MS_{bg}}{MS_{error}}$$

where

$$MS_{error} = \frac{SS_{error}}{df_{error}}$$

and

$$df_{error} = df_{wg} - df_{subj}$$

C.6 Two-Way ANOVA

C.6.1 Introduction

Sometimes the data to be analyzed can be divided using two conditions (factors), for example an experiment with 3 time points (20 min, 1 h, 6 h) and 2 drug doses (10µg, 10µg). In the table below the experimental design of a two-factor, or Two-Way ANOVA, problem is shown. Each combination of the levels of the two factors is called a cell. Thus, samples 1, 2 & 3 are in the same cell. Since all cells have the same number of samples the design is said to be balanced (see unbalanced design section).

Examples

Time\Dose	**1mg**	**10mg**
20 min	Sample 1, 2 & 3	Sample 4, 5 & 6
1 h	Sample 7, 8 & 9	Sample 10, 11 & 12
6 h	Sample 13, 14 & 15	Sample 16, 17 & 18

Table C-4. Schematic example of a Two-Way ANOVA experimental design.

With this experimental design, three different hypotheses can be tested in the experiment. Firstly, tests if the time or dose significantly changes the dependent variable can be performed. With a two-factor design a test can also be performed to see if there is an interaction effect between the two factors. Do the two conditions affect each other?

Repeated measures

Two-Way ANOVA with repeated measures for both factors are similar to normal Two-Way ANOVA but there are values for a subject for all combinations in the experimental design.

Time\Dose	**Subject**	**1mg**	**10mg**
20 min	1	Sample 1& 2	Sample 3& 4
	2	Sample 5& 6	Sample 7& 8
	3	Sample 9& 10	Sample 11& 12
	4	Sample 13& 14	Sample 15& 16
1 h	1	Sample 17& 18	Sample 19& 20
	2	Sample 21& 22	Sample 23& 24
	3	Sample 25& 26	Sample 27& 28

Time\Dose	Subject	1mg	10mg
	4	Sample 29& 30	Sample 31& 32
6 h	1	Sample 33& 34	Sample 35& 36
	2	Sample 37& 38	Sample 39& 40
	3	Sample 41& 42	Sample 43& 44
	4	Sample 45& 46	Sample 47& 48

Table C-5. Schematic presentation of the Two-Way ANOVA with different subjects.

C.6.2 Detailed Description

The following sections describe the traditional way of performing Two-Way ANOVA calculations. In DeCyder EDA however, the ANOVA algorithms are implemented using multiple linear regression analysis to handle unbalanced data sets (groups with different sizes). It can be proved mathematically that analysis of variance and regression are two ways of calculating the same result.

For further information on the implementation, see the reference list.

Traditional Independent Two-Way ANOVA

In Two-Way ANOVA, the total sum of variance can be portioned in a similar way as in the One-Way ANOVA case. In this case however, some of the variance is associated with each of the two factors in the experiment, and some variance is associated with the interaction between the two factors. The remaining variance is the residual or random error:

$$SS_{TOT} = SS_A + SS_B + SS_{AB} + SS_{res}$$

where **SS_{res}** is the sum of squared deviation of all the observations from their respective cell means.

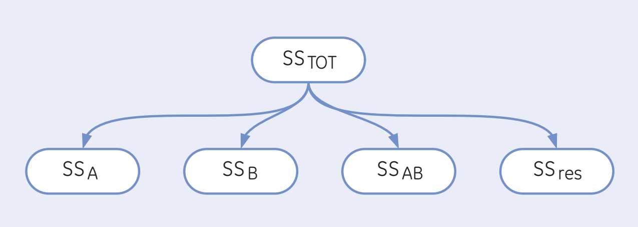


Fig C-6. Portioned Variability in the Two-Way ANOVA problem.

The null hypothesis when calculating Two-Way ANOVA is that neither of the changes in treatment levels nor the interaction is associated with observed values.

As in One-Way ANOVA, the $\mathbf{SS_{TOT}}$ is calculated first, as the total sum of squared deviations from the total mean.

The random error, the residual, is then calculated as the sum of squared deviation from the respective means:

$$SS_{res} = \sum_i \sum_j \sum_k \left(x_{ijk} - \bar{x}_{ijk} \right)^2$$

where **i** and **j** are the two factors and their levels and **k** is the subjects.

The $\mathbf{d_{fres}}$ is associated with ab(n - 1) where **a** and **b** are the number of levels for treatment A and B and n is the number of observations in each cell.

Thus:

$MS_{res} = SS_{res} / df_{res}$

The variability in factor **A** and **B** is computed in a similar way to One-Way ANOVA and is not explained here (see reference for full description).

To test whether the variability between the different levels of **A** or **B** by chance is greater than expected, one calculates:

$F_A = MS_A / MS_{res}$ *and* $F_B = MS_B / MS_{res}$

These values are then tested against the F distribution with $\mathbf{df_A}$ or $\mathbf{df_B}$ and $\mathbf{df_{res}}$.

To test the interaction effects, one calculates the $\mathbf{SS_{AB}}$ from the already calculated SS terms and the same is true for the $\mathbf{df_{AB}}$ term.

So,

$F_{AB} = MSAB / MSres$

is tested against the F distribution with $\mathbf{df_{AB}}$ and $\mathbf{df_{res}}$.

Since there can be several factors (conditions) in EDA, e.g. time, dose and temperature, the user has an option to select which two conditions to use in the Two-Way ANOVA calculation.

Traditional Repeated Measures Two-Way ANOVA

As with other ANOVA calculations, the total sum of squares of the observations, $\mathbf{SS_{TOT}}$, can be used to portion the variability of the data into the different aspects.

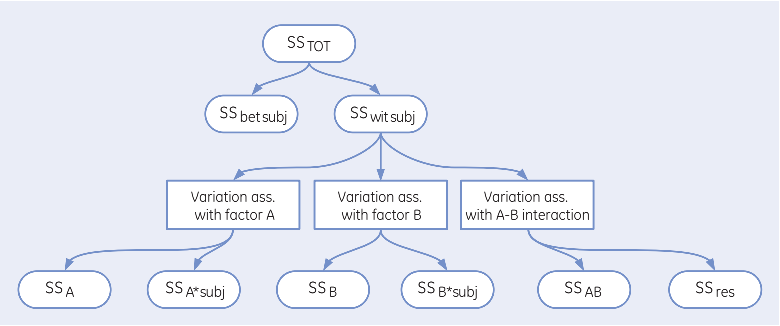

SS TOT
SS bet subj
SS wit subj
Variation ass. with factor A
Variation ass. with factor B
Variation ass. with A-B interaction
SS A
SS A*subj
SS B
SS B*subj
SS AB
SS res

Fig C-7. Portioned variability of the repeated measures Two-Way ANOVA.

The procedure for calculating RM Two-Way ANOVA is similar to the previous ANOVA algorithms where sums of squares are calculated for the different parts of the variability.

1 Calculate if factor A has an effect on the outcome.

Part of the variation in A is due to the different A levels, not considering B levels and the subjects: $\mathbf{SS_A}$.

At the same time the different subjects can be followed over all A levels (in balanced sets) since it is the same subject (individual).

The difference in response levels for each subject for each A level can then be estimated: $\mathbf{SS_{A*subj}}$

By dividing by the degrees of freedom, the Mean squares can be created and thus the F value:

$$\mathbf{F_A = MS_A / MS_{A*subj}}$$

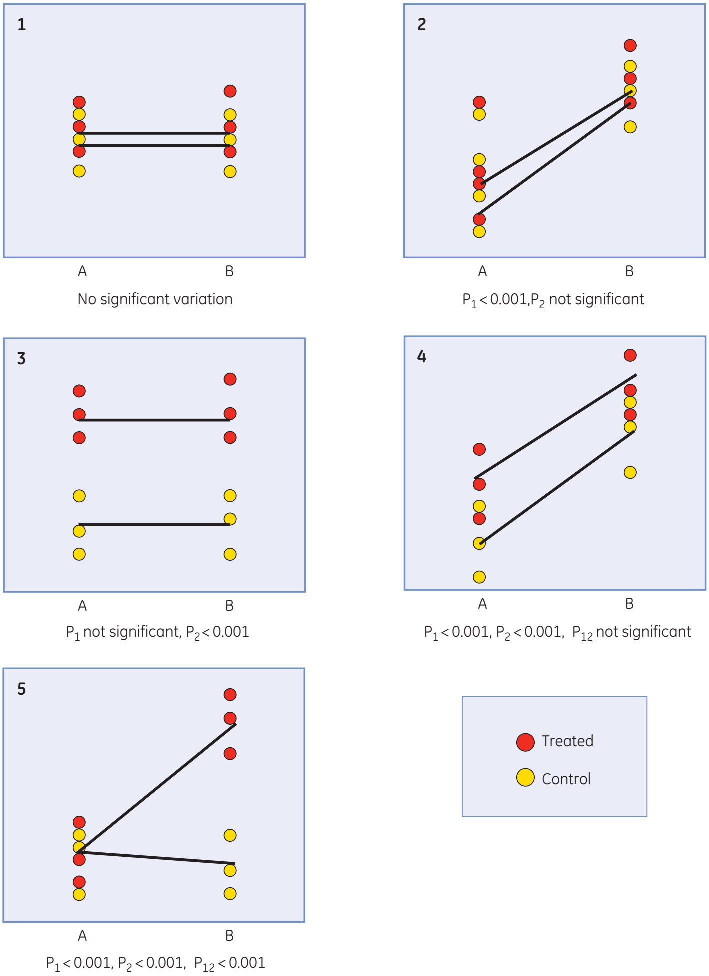
This value is then compared to the F distribution to receive the p-value for factor A.

By analyzing the above, one can see that the test is to see where the variability in A comes from. If it is due to the levels of A, then a high $\mathbf{MS_A}$ value is received, whereas if it is largely dependent on the different subjects the $\mathbf{MS_{A*subj}}$ has a high value.

2 $\mathbf{F_B}$ is calculated in a similar way as $\mathbf{F_A}$.

3 The interaction term $\mathbf{F_{AB}}$ is the last one to be calculated, and is calculated similarly as with the Independent Two-Way ANOVA, but with the subjects in mind:

$$F_{AB} = MS_{AB} / MS_{res}$$



No significant variation

$P_1 < 0.001, P_2$ not significant

P_1 not significant, $P_2 < 0.001$

$P_1 < 0.001$, $P_2 < 0.001$, P_{12} not significant

$P_1 < 0.001$, $P_2 < 0.001$, $P_{12} < 0.001$

Fig C-8. Graphical examples of Two-Way ANOVA analyses.

The graphs in Fig. C-8 illustrate changes in protein abundance (y-axis) for a two-condition experiment. Condition 1 (x-axis) represents two temperatures A and B; condition 2 (red and yellow circles) represents drug treated and control samples.

Each condition is in triplicate, hence there are four experimental groups with 3 samples in each group. Conditions 1 and 2 are used to link groups together based on one common factor, i.e. group 1 and 2 may have the same condition 1 value (both temperature 1) but different condition 2 value (drug treated or control). Groups 3 and 4 will have the other condition 2 value (both temperature 2) with different condition 2 values (drug treated or control).

C.7 Multiple Test Correction

C.7.1 Introduction

An EDA workspace can contain thousands of proteins and when a statistical test such as Student's T-test or One-Way ANOVA is run, a value for each of the proteins is calculated.

This means that thousands of hypotheses are tested at the same time, which will lead to an increased chance of false positives (proteins that are falsely said to be differentially expressed when they are not). The multiple test correction methods adjust the p-values to account for occurrences of false positives.

The multiple test correction in EDA uses a method called False Discovery Rate (FDR).

The multiple testing area is a dynamic area where new corrections are published very frequently, but the algorithm implemented in EDA is the adaptive FDR of Hochberg and Benjamini (2000).

C.7.2 Conceptual example

Imagine a box with 10 balls, 9 are white and 1 is black.

What is the probability of getting the black ball? (10%).

If this was repeated 20 times and after each time the ball that was picked was placed in the box again, what is the probability of getting the black ball during one of the 20 rounds?

It's definitely much higher...

In fact it's 100 * (1-0.920) = 88%

This conceptual example can be translated into differential expression analysis, if the black ball is thought to be a false positive. So by doing the test several times, the risk of getting a false positive is increasing.

Example

A control-treatment experiment contains 3000 proteins. Since each protein is assumed to be independent each protein is tested individually. The p-value threshold is 0.01.

Then 3000 * 0.01 = 30 proteins will by chance have a p-value < 0.01.

If really significantly expressed, the threshold of 0.01 is no longer valid.

Thus the p-values need to be adjusted for multiple testing.

C.8 Multiple Comparisons

C.8.1 Introduction

In One-Way ANOVA, the p-value that is calculated will indicate if the mean of the groups are significantly different or not. A major drawback with this method is that it doesn't indicate between which groups the mean difference was significant.

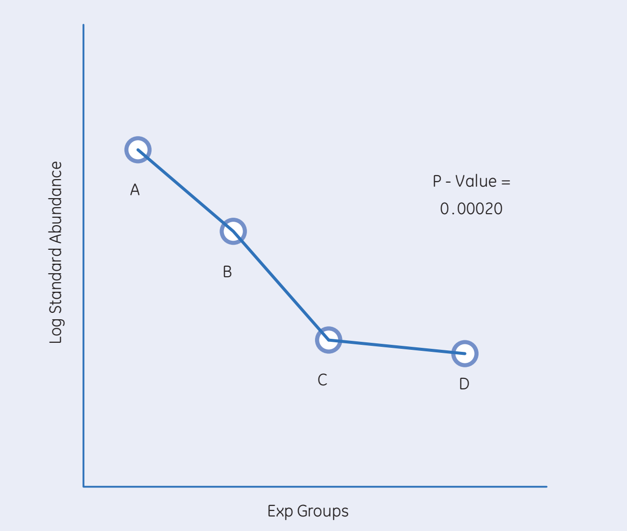


Fig C-9. One-Way ANOVA value indicates significant differences in protein expression, but it doesn't indicate between which groups (A-D probably, but is A-C significant?).

There are a number of multiple comparison methods that can be used to investigate between which groups the difference in protein expression is significant. The Tukey's multiple comparison test, or, as it is also called, Tukey's

Honestly Significant Difference Test (Tukey's HSD) is included in EDA for this purpose.

C.8.2 Method

The Tukey's HSD can be calculated to do pair-wise comparisons between all groups.

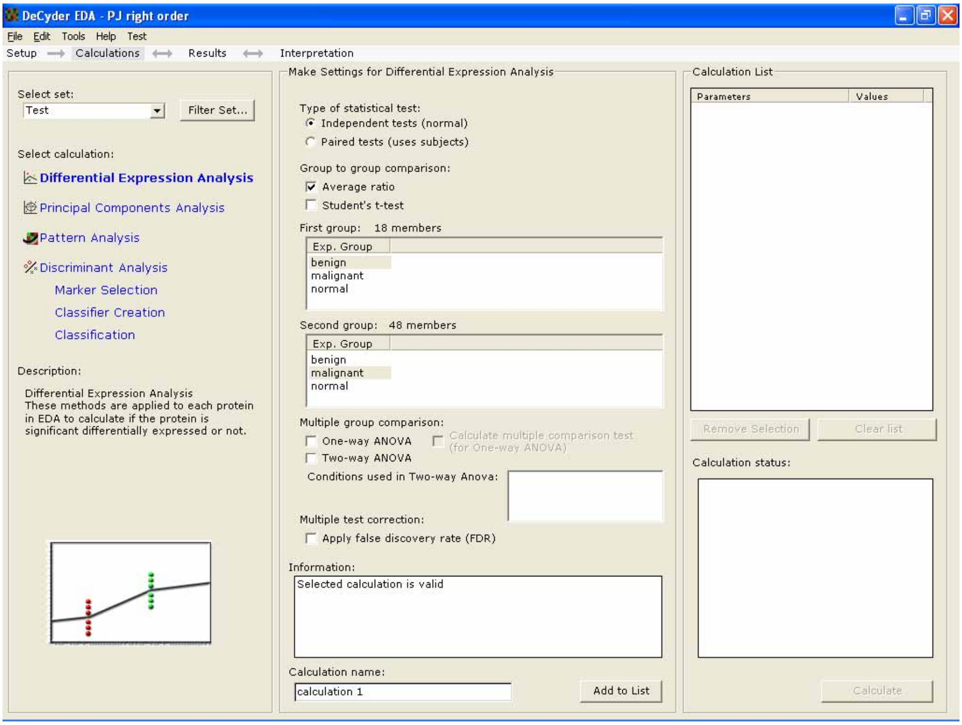
The critical value Q for each pair of groups is calculated as follows:

$$Q = \frac{mean_i - mean_j}{\sqrt{\dfrac{MS}{hm_{i,j}}}}$$

where **mean**$_{i \text{ or } ,j}$ is the mean of group **i** and **j**. **MS** is the mean square error and **hm**$_{ij}$ is the harmonic mean of the sample sizes of group **i** and **j**.

The Studentized Range Distribution is then used to calculate the P value, using the Q values and the number of samples and the number of degrees of freedom associated with the original ANOVA calculation.

C.9 Calculation Setup



Select set: Test

Select calculation: Differential Expression Analysis

Description:
Differential Expression Analysis
These methods are applied to each protein in EDA to calculate if the protein is significant differentially expressed or not.

Make Settings for Differential Expression Analysis

Type of statistical test:
● Independent tests (normal)
○ Paired tests (uses subjects)

Group to group comparison:
☑ Average ratio
□ Student's t-test

First group: 18 members

Exp. Group
benign
malignant
normal

Second group: 48 members

Exp. Group
benign
malignant
normal

Multiple group comparison:
□ One-way ANOVA □ Calculate multiple comparison test (for One-way ANOVA)
□ Two-way ANOVA
Conditions used in Two-way Anova:

Multiple test correction:
□ Apply false discovery rate (FDR)

Information:
Selected calculation is valid

Calculation name: calculation 1

Calculation status:

Fig C-10. EDA screenshot of Differential Expression Analysis calculation setup.

Parameter	Description
Type of Statistical test	Select independent or paired test (see description of tests above). Note that for paired tests subject information needs to be entered for spot maps.
Group to group comparison	Mark check boxes to calculate Average ratio or Student's T-test.
Multiple group comparison	Select the groups to run the test between in the two lists.Note that one can select several groups in each list. Mark check boxes to calculate One-Way or Two-Way ANOVA. For One-Way ANOVA, select to calculate multiple comparison test.
Multiple Test Correction	To correct for multiple testing by a False Discovery Rate, check the box.

Parameter	Description
Calculation Name	Enter a name for the calculation.

Table C-6. Settings and parameters for Differential Expression Analysis.

C.10 References

1 Primer of Applied Regression and Analysis of Variance 2nd Edition (Glantz and Slinker) McGraw Hill (2000).

2 Benjamini Y, Hochberg Y, On the adaptive control of the false discovery fate in multiple testing with independent statistics, J EDUC BEHAV STAT 25 (1): 60-83 SPR (2000).

Appendix D Statistics and algorithms - Principal Component Analysis

D.1 Introduction

Principal Component Analysis (PCA) is essentially a method for reducing the dimension of the variables in a multidimensional space. Multivariate data consists of objects that have been observed using a number of variables and the PCA algorithm analyses the data to try and reduce the number of variables, since some of the variables often correlate.

For example, consider a photograph of an apple. The information about the apple has been projected from a 3-dimensional world onto a 2-dimensional piece of paper. However, since it can be recognized as an apple, there is enough information. One dimension has been removed but most of the information is retained.

The projections in PCA are often described as a definition of a new set of coordinate axes, the data isn't changed, it's just the axes.

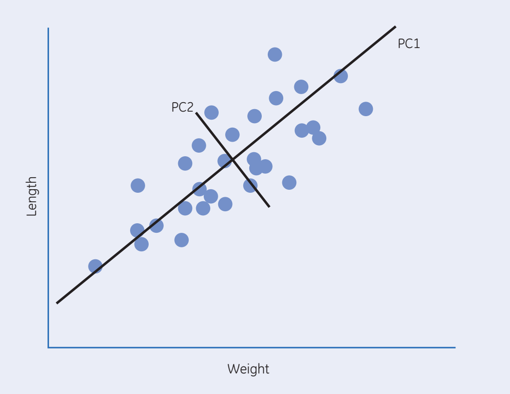


Fig D-1. An example of correlation between height and weight where PCA has been calculated. The first principal component (PC1) goes in the direction where the most variance is situated. The second principal component is perpendicular to the first one and accounts for the second most variance.

Several methods exist that determine the principal components of a dataset, and they all extract PCs in decreasing order, so that the first principal component contains the most information (most of the variability in dataset) and each successive component accounts for a little less.

After the PCA analysis, one tries to interpret the first few principal components in terms of the original variables, and thereby get a greater understanding of the data. To reproduce the total system variability of the original p variables, all p PCs are needed. However, if the first few PCs account for a large proportion of the variability (80-90%), the objective of dimension reduction is achieved.

In DeCyder EDA, the user can decide if proteins, spot maps or experimental groups shall be variables or observables.

Scores	Loadings	Description
Proteins	Spot Maps	Tries to reduce the number of spot maps, to get a simpler view of the proteins in the data set.
Proteins	Experimental Groups	Tries to reduce the number of experimental groups, to get a simpler view of the proteins in the data set.
Spot Maps	Proteins	Tries to reduce the number of proteins, to get a simpler view of the spot maps in the data set.
Experimental Groups	Proteins	Tries to reduce the number of spot maps, to get a simpler view of the experimental groups in the data set.

Table D-1. Overview of analysis strategies available in Principal Component Analysis.

With PCA in DeCyder EDA it is thus possible to detect outliers in the data set, data points far away in the Score plot and initial clusters in the score plot. It is also possible to identify the spot maps that have similar expression profiles and if replica sets are homogenous or not.

Example

In this example a PCA has been performed on a data set with 76 spot maps and 121 proteins, where the spot maps belong to 3 different experimental groups.

By analyzing the first two PCs one can see total separation of the different spot maps into the 3 different groups.

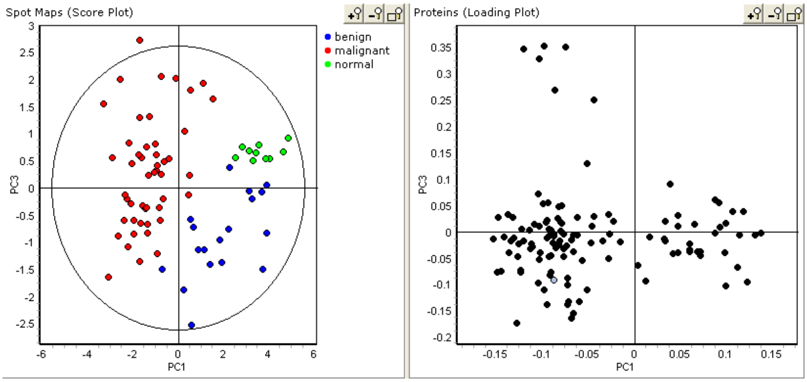


Fig D-2. A PCA example result indicating that the three experimental groups can be separated.

D.2 Detailed Description

D.2.1 Assumptions and characteristics

As with all statistical techniques there are assumptions about the data in PCA. The main assumption is that the derived components are normally distributed and uncorrelated (orthogonal). If PCA is being used to test statistical hypotheses the assumptions should be valid. The assumptions are less important when PCA is used as a descriptive and exploratory tool. In practice, if the principal components are normally distributed the assumptions may be considered valid.

D.2.2 PCA Algorithm

The PCA implementation uses the NIPALS algorithm for calculation of the principal components since it is not as memory demanding as other algorithms such as Single Value Decomposition (SVD).

The scores are often denoted as **T** and the loadings as **P**. For each score variable t, the influence of the original variables is found in its corresponding loading profile, **p**. This provides a direct link between the scores **T**, and the original **X**-variables and is used when interpreting the results using plots of scores (t1-t2) and loadings (p1-p2).

The NIPALS method works like this:

1	Select a column in **X**, set to starting vector **t**	
2	$p = (X't) / (t't)$	(**X** is projected onto t to find the corresponding loading p)
3	$p = p / \|\| p \|\|$	(Length of vector **p** is set to 1)
4	$t = (Xp) / (p'p)$	(**X** is projected onto **p** to find corresponding score vector **t**)
5	If the difference between **t** in step 4 and **t** in step 1 is larger than a pre-defined threshold, there is no convergence yet and the algorithm returns to step 1.	
6	$E = X - tp'$	(The estimated component is removed from **X**)

The part of **X** that is not explained by the model forms the residuals **(E**). To estimate more than one component, the procedure is repeated, but **X** is replaced by **E** in step 1.

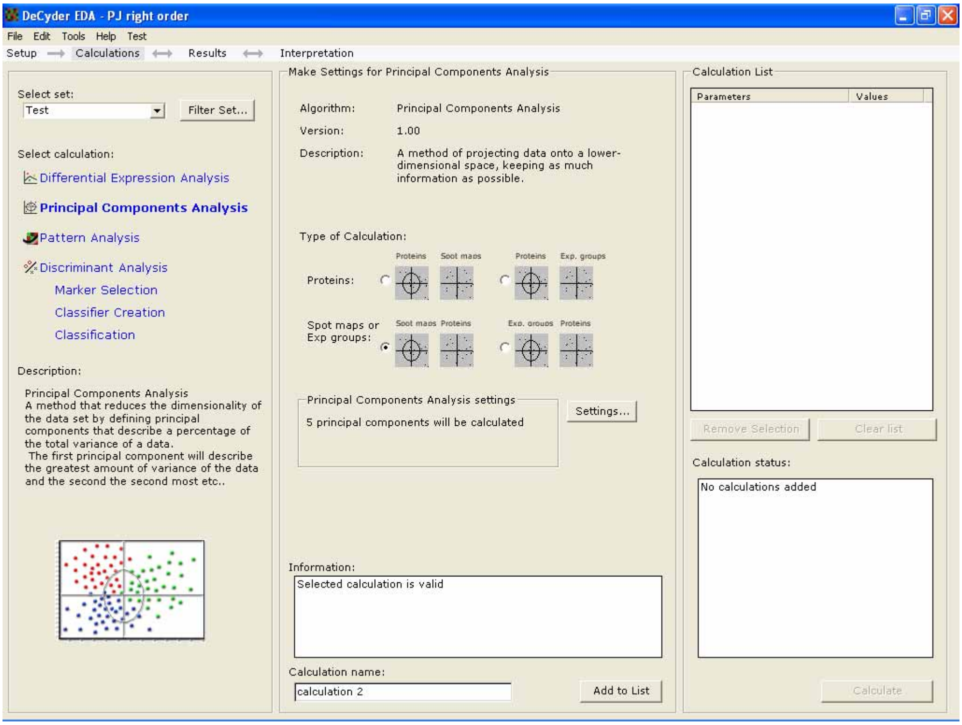
PCA analyzes the data space and finds a low-dimensional hyper-plane that best summarizes all the variation in **X**, in terms of least squares. The coordinates of the points projected onto this hyper-plane are called scores **t**. The direction of each dimension in the hyper-plane is its loading **p**. The loading values (weight) for each PCA component are the cosine of the angles between the principal component direction and the original coordinate axes and correspond to the distance between each point in K-space and its point on the plane. The scores, loadings and residuals together describe all of the variation in **X**:

Model of X: $\mathbf{X} = \mathbf{TPT} + \mathbf{E} = \mathbf{t_1p^T1} + \mathbf{t_2p^T2} + ... + \mathbf{E}$

The loadings (**P**) are ranked in the order of the largest eigenvectors of (**X'X**) and the score vectors (**T**) are ranked in the order of the largest eigenvectors of (**XX'**).

Prior to PCA calculations, the PCA algorithm in EDA column centers the **X** matrix. This corresponds to moving the center of the swarm of points to the origin.

D.3 Calculation Setup



Select set: Test

Select calculation:

Description:

Principal Components Analysis
A method that reduces the dimensionality of the data set by defining principal components that describe a percentage of the total variance of a data.
The first principal component will describe the greatest amount of variance of the data and the second the second most etc..

Algorithm: Principal Components Analysis

Version: 1.00

Description: A method of projecting data onto a lower-dimensional space, keeping as much information as possible.

Type of Calculation:

Proteins: ○ Proteins / Spot maps ○ Proteins / Exp. groups

Spot maps or Exp groups: ● Spot maps / Proteins ○ Exp. groups / Proteins

Principal Components Analysis settings

5 principal components will be calculated

Information:

Selected calculation is valid

Calculation name: calculation 2

Calculation status:

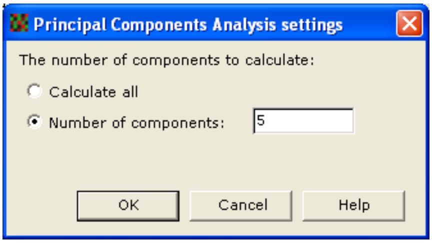
No calculations added

Fig D-3. EDA screenshot of Principal Component Analysis calculation setup.

The Principal Component Analysis can be calculated for proteins, spot maps and experimental groups.

Parameter	Description
Number of components to calculate.	Either all principal components can be calculated or just a few. There is very rarely a need to calculate all components since much of the variance is covered by the first components, and therefore the default setting is 5 PCs in EDA.

Table D-2. Settings and parameters for Differential Expression Analysis.


Principal Components Analysis settings

The number of components to calculate:

○ Calculate all

● Number of components: 5

Fig D-4. EDA screenshot of Principal Component Analysis settings dialog.

D.4 References

NIPALS

H. Wold, Estimation of principal components and related models by iterative least squares in Multivariate Analysis (Ed., P.R. Krishnaiah), Academic Press, NY, 1966, pp. 391-420.

PCA and NIPALS

Multi- and Megavariate Data Analysis, I Eriksson, E Johansson, N Kettaneh Wold- and S. Wold, Umetrics Academy, Umeå 2001, ISBN 91-973730-1-X, 533p

SIMCA

SIMCA 10.5 Umetrics AB Sweden

Appendix E Statistics and algorithms - Pattern Analysis

E.1 Introduction

Pattern Analysis, also called Cluster Analysis or unsupervised clustering, is a process to group similar objects together. In the images in Fig. E-1 below, several animals have been grouped together since they are animals.

To cluster one needs to find out what to cluster, thus how to define similarity. The animals can be clustered even further, e.g. into carnivore and herbivore clusters, or clusters of animals with two feet and four feet. It's up to whoever performs the clustering to define the similarity measure.

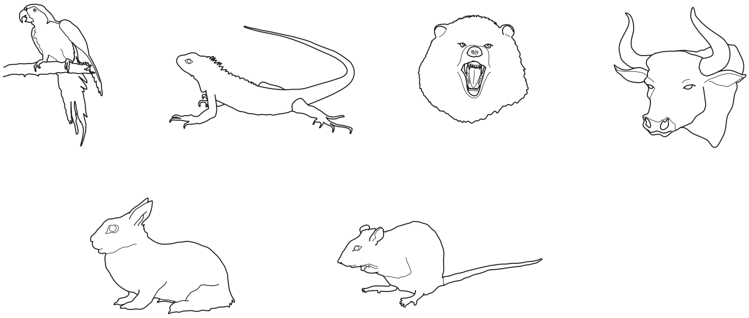


Fig E-1. Animals have been grouped together. These animals can be sub grouped but the similarity measure needs first to be defined.

The pattern analysis in EDA consists of algorithms that can help in analyzing the expression matrix and to find these subsets of data (clusters) that show similar expression patterns.

For example:

- Identify similarities in protein expression and group proteins.
- Identify diagnostic and prognostic markers from protein expression patterns.
- Identify samples that have similar expression profiles e.g. in a tumor type experiment.
- Identify experimental groups (experimental replicas) that show similar expression profiles.

The data set can thus be grouped in both dimensions in DeCyder EDA.

Similar objects can then by the general hypothesis "Guilt by association" be assumed to have something in common, for example similarly expressed proteins give rise to a hypothesis of co-regulation or a functional relationship.

E.1.1 Quality

A good cluster should have low variance within the cluster, thus be homogenous, but have a large variance against other clusters (see also Dunn and R2 below).

E.1.2 Unsupervised clustering types

There are numerous unsupervised clustering algorithms, but these can be divided into two groups; the hierarchical, where the results are in the form of a hierarchical tree, and the partitioned, which groups data into distinct groups.

The EDA application contains algorithms from both categories.

Method	Description	Application
Hierarchical Clustering	Hierarchical Clustering. Clustering method to group the data in a hierarchical way, a dendrogram	Is used for clustering data to find correlated expression profiles or samples that have the same expression levels over all proteins. Standard clustering algorithm.
K-means	Partition Clustering. Clustering method that divides the data into distinct groups.	Is used for clustering data to find correlated expression profiles or samples that have the same expression levels over all proteins. Mainly used for time-dose or similar applications.
Self-Organizing Maps	Partition Clustering. Clustering method that divides the data into distinct groups that are linked.	Is used for clustering data to find correlated expression profiles or samples that have the same expression levels over all proteins. Mainly used for time-dose or similar applications.

Method	Description	Application
Gene Shaving	Partition Clustering. Clustering method that identifies small sets of proteins with coherent expression patterns and differs from other widely used methods in that items may belong to more than one cluster.	Is used for identifying a smaller set of proteins that can be included in several clusters.

Table E-1. Clustering Methods in EDA.

E.2 Similarity Measures

In EDA there are two different similarity measures implemented for the pattern analysis methods; the Euclidean Distance and the Pearson Correlation Coefficient.

E.2.1 Euclidean Distance

The Euclidean distance is a particularly common distance measurement. It is calculated as:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

It is essentially the sum of squared distances of two vector values, e.g. protein expression values in a spot map.

E.2.2 Pearson Correlation Coefficient

Pearson Correlation Coefficient is the most widely used measurement of association between two vectors.

A correlation coefficient results in a value between -1 and 1, where a value of -1 means that the vectors are completely opposite to each other, 0 means that they are completely independent and 1 means that they are identical.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ is the mean of vector **x** and **y** respectively.

Since a distance measure **d** is needed, where a distance of 0 means that the vectors are identical, the value **r** is transformed into a distance measure by:

d = 1 - r

E.2.3 Comparison of similarity measures

The Euclidean distance takes into account the absolute values whereas the Pearson correlation measurement can be used to evaluate trends of expression over a set of treatments when the magnitude isn't of importance.

The Pearson and Euclidean distance give the same result when the vectors have been normalized (using mean centering and standard deviation).

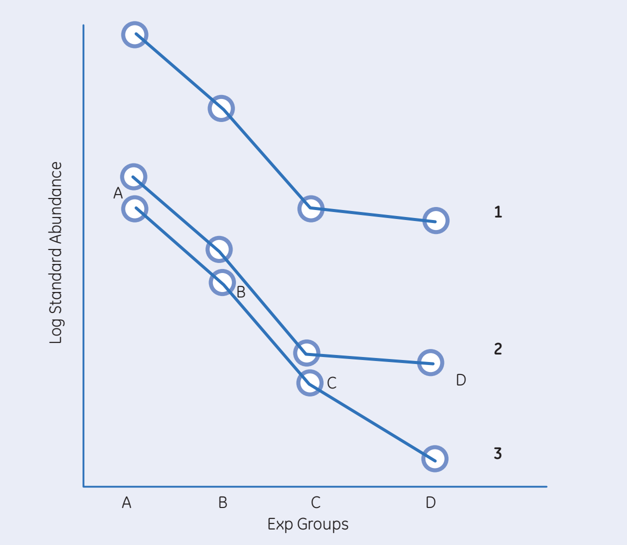


Fig E-2. Three protein expression profiles over a four exp group interval.

In the figure above the different similarity measures gives the following result:

- Euclidean: Protein 2 and 3 are most similar
  (Note the closeness of their absolute values)
- Pearson: Protein 1 and 2 are most similar
  (Note the closeness of their relative values - trend)

It is up to the user to define what similarity means and then select the appropriate similarity metric. The default measure in EDA is Euclidean distance.

E.3 Hierarchical Clustering

E.3.1 Introduction

The perhaps most widely used unsupervised clustering algorithm is hierarchical clustering, which is a method that combines or splits the data two by two and thereby generates a treelike structure called a dendrogram.

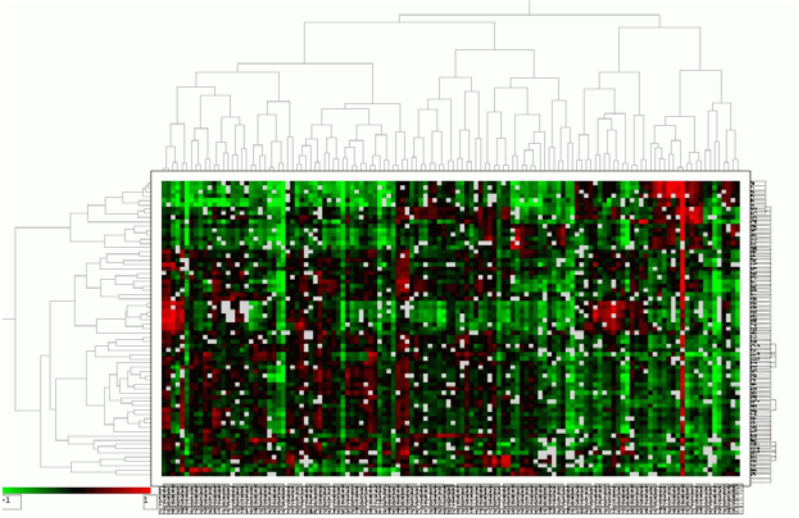


Fig E-3. Image of a dendrogram and a heat map (expression matrix). All the nodes to the right are called leaf nodes and the single node to the left, the root node.

- One of the advantages with hierarchical clustering is that very few parameters need to be specified; the distance measure and the linkage rule.
- The resulting tree will not only give the similarities but also the distances (branch lengths) which could be interesting in some applications.
- One drawback of hierarchical clustering is that actual clusters are not formed. Instead it is up to the user to define clusters depending on the branching pattern.
- It might be computationally difficult for a normal computer to calculate similarity matrices of tens of thousands of objects.

Example

In an experiment, the log standard abundance of 4 proteins were measured for two spot maps:

	Spot map 1	Spot map 2
Protein 1	0.3	0.2
Protein 2	0.25	0.3
Protein 3	0.5	0.5
Protein 4	0.7	0.4

By calculating the distances between the proteins using Euclidean distance the algorithm comes up with this result:

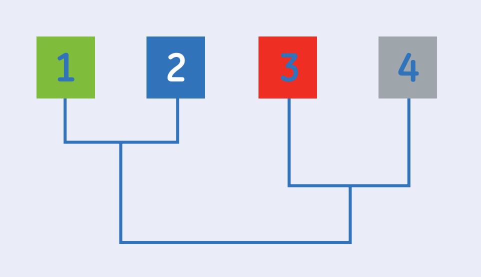


Fig E-4. The protein 1 and 2 have the most similar profile according to Euclidean distance and are therefore clustered together. Protein 3 and 4 have fairly similar profiles but the two groups 1,2 and 3,4 are not very similar and are therefore joined at the bottom.

E.3.2 Detailed Description

The hierarchical clustering in EDA is agglomerative, and conceptually works like this:

1. All pair-wise distance measures, between every two objects, are calculated and a similarity matrix is constructed. All objects in this moment are now leaves.

2. The smallest number in the similarity matrix indicates the two most similar nodes **r** and **c**.

3. These two are then merged (linked) and **r** is replaced by the new node whereas **c** is removed from the similarity matrix. The distances that have been affected are recalculated.

4. Repeat step 2 and 3 until there is only one node left, the root node.

Linkage

When merging two nodes there is a need for a rule or method for how to define the distance from the new node to all of the other nodes. This is needed to update the distance matrix used for similarity measures.

The following linkage methods are available in EDA:

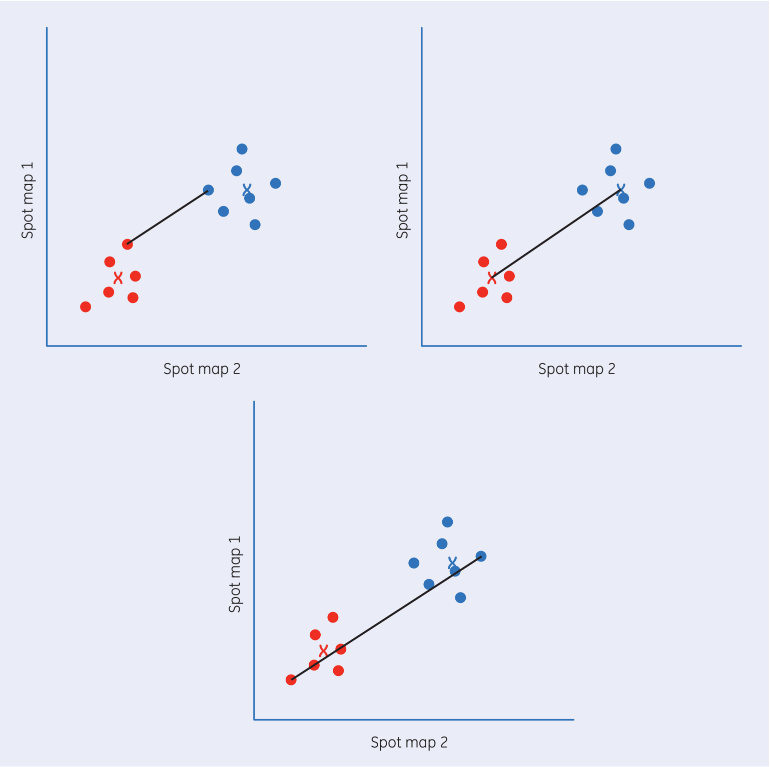
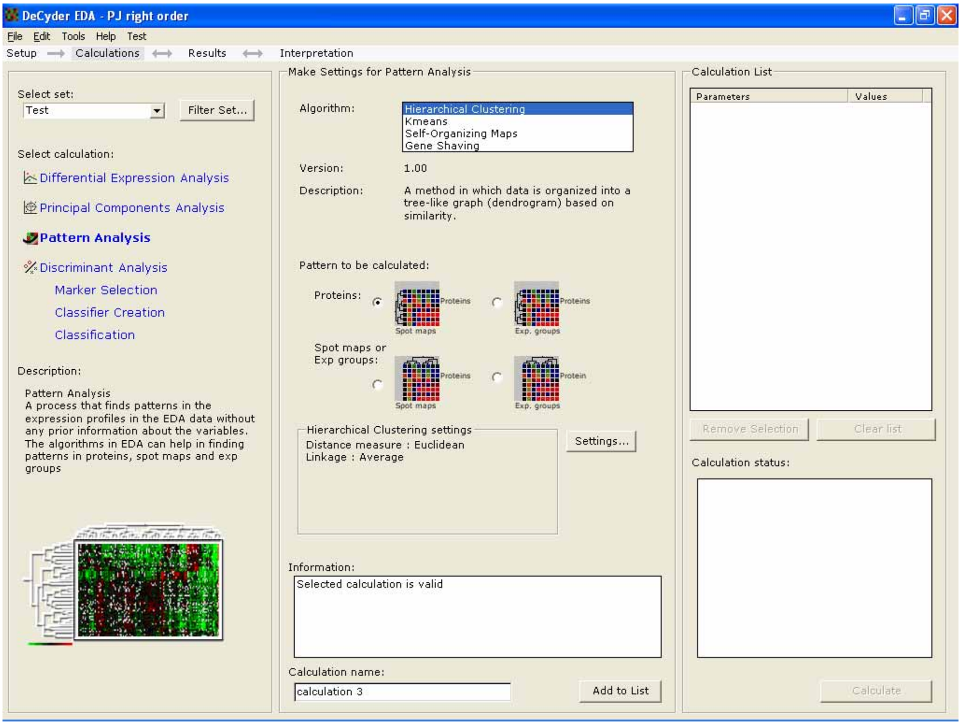


Fig E-5. Linkage of two clusters with single, average and complete linkage. The distance between the two clusters in each diagram is indicated by a straight line.

Linkage Method	Definition	Description
Single Linkage	Defines the distance between two nodes as the distance between the closest objects between the two clusters.	Produces a skewed hierarchy (chaining problem), where only one small distance may link two otherwise very different clusters. The main advantage is that outlying objects are easily identified by this method, as they will be the last to be merged.
Average Linkage	Defines distance between clusters as the distance between the cluster centroids.	Is more stable with respect to unknown data point distributions than the other two methods.
Complete Linkage	Defines distance as the distance between the farthest pair of points in the two clusters.	Tends to be less desirable when there is a considerable amount of noise present in the data. Complete linkage produces very compact clusters.

Table E-2. Linkage methods in Hierarchical Clustering analysis.

E.3.3 Calculation Setup



Select set: Test

Select calculation: Pattern Analysis

Description:
Pattern Analysis
A process that finds patterns in the expression profiles in the EDA data without any prior information about the variables. The algorithms in EDA can help in finding patterns in proteins, spot maps and exp groups

Make Settings for Pattern Analysis

Algorithm: Hierarchical Clustering
Kmeans
Self-Organizing Maps
Gene Shaving

Version: 1.00

Description: A method in which data is organized into a tree-like graph (dendrogram) based on similarity.

Pattern to be calculated:

Proteins: ● ○

Spot maps or Exp groups: ○ ○

Hierarchical Clustering settings
Distance measure : Euclidean
Linkage : Average

Information:
Selected calculation is valid

Calculation name: calculation 3

Calculation List

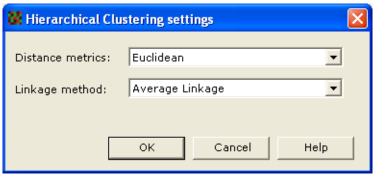
Calculation status:

Fig E-6. EDA screenshot of Hierarchical Clustering calculation setup.

Hierarchical Clustering can be calculated for proteins, spot maps and experimental groups by selecting the corresponding button.

Parameter	Description
Distance metrics	Choose Euclidean or Pearson Correlation. Euclidean distance is the default setting.
Linkage method	Choose Single, Average or Complete linkage. Average linkage is the default setting.

Table E-3. Settings and parameters for Hierarchical Clustering analysis.



Hierarchical Clustering settings

Distance metrics: Euclidean

Linkage method: Average Linkage

Fig E-7. EDA screenshot of Hierarchical Clustering settings dialog

E.4 K-means

E.4.1 Introduction

K-means clustering is one of the oldest ways to cluster objects but is still one of the most common ways to do it. It divides the objects into a predefined number of clusters, k, so that each object belongs to just one cluster. The basic idea behind the algorithm is to move the centroids during the iterations and put each object into a cluster depending on similarity.

The traditional K-means algorithm is very fast and simple, but it has some limitations:

- The number of clusters have to be defined in advance
- The initial position of the centroids is random.

Both these limitations have been addressed in EDA by using gap statistics as a tool to estimate the number of clusters and by using hierarchical clustering result as starting points.

Example

In a time or dose experiment, K-means can be used to find the proteins that have the same expression profile over, for instance, time or dose.

	Time point 1	**Time point 2**	**Time point 3**	**Time point 4**	**Time point 5**
Protein 1	0.3	0.4	0.6	0.7	0.9
Protein 2	0.25	0.3	0.55	0.7	0.8
Protein 3	0.5	0.4	0.5	0.5	0.5

In this example protein 1 and protein 2 have similar expression profiles over time.

If the K-means algorithm had been used it would have put these proteins into the same cluster based on their expression profile.

E.4.2 Detailed Description

1 The traditional K-means algorithm conceptually works like this.

2 The k centroids are randomly positioned and the objects randomly assigned to a centroid. The mean of the centroids is then calculated.

3 Each object is associated with the closest centroid according to a distance measure between the object and the centroid means.

4 The means of the centroids are re-calculated.

Steps 3 and 4 are repeated until no object changes its association to a centroid or until a certain number of iterations have been reached, since the algorithm may not converge.

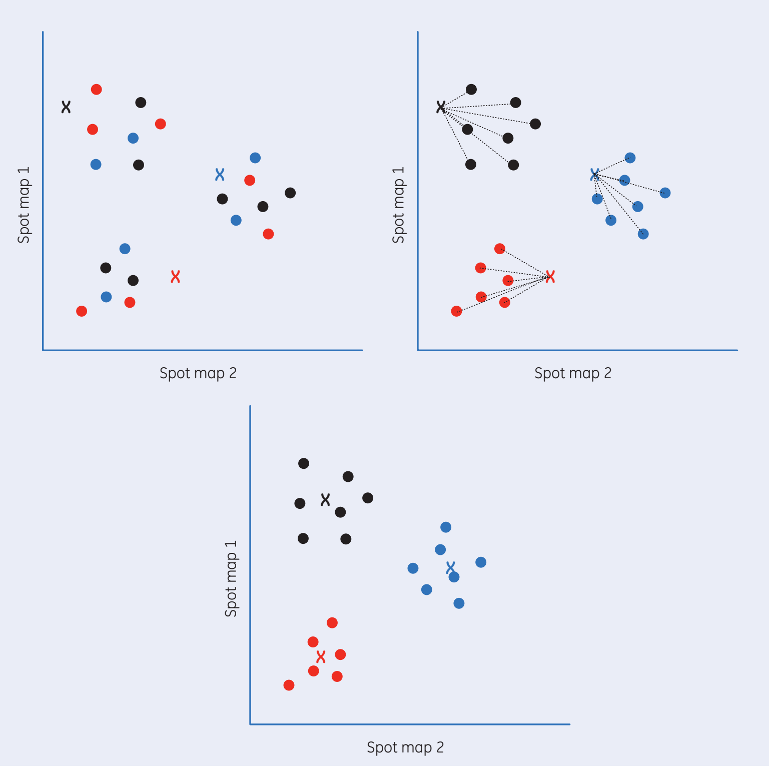
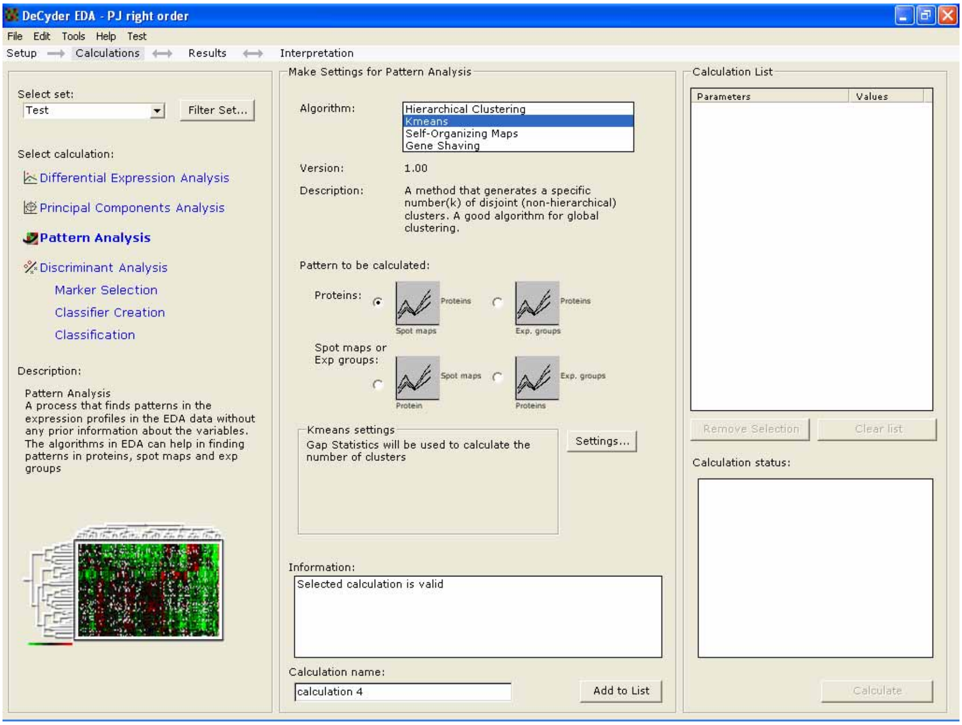


Fig E-8. The three images show three steps in the K-means algorithm. 1) Randomly positioned centroids and associations. 2) Association with the closest centroid. 3) The means of the centroid are recalculated.

In traditional K-means, the random starting positions are crucial for the result, thus the result is not deterministic. The algorithms will not end up with the same result if a calculation is redone, unless the same starting points are used.

To enable deterministic results, a hierarchical clustering is performed first to calculate the starting positions for the K-means algorithm. The dendrogram is cut at the number of clusters defined in K-means, the centroids are then placed at the mean positions of the proteins that are in each cut node in the dendrogram.

E.4.3 Calculation Setup



Select set: Test

Select calculation: Pattern Analysis

Description:

Pattern Analysis
A process that finds patterns in the expression profiles in the EDA data without any prior information about the variables. The algorithms in EDA can help in finding patterns in proteins, spot maps and exp groups

Algorithm: Kmeans

Version: 1.00

Description: A method that generates a specific number(k) of disjoint (non-hierarchical) clusters. A good algorithm for global clustering.

Proteins: ● Proteins / Spot maps ○ Proteins / Exp. groups

Spot maps or Exp groups: ○ Spot maps / Protein ○ Exp. groups / Proteins

Kmeans settings
Gap Statistics will be used to calculate the number of clusters

Information:
Selected calculation is valid

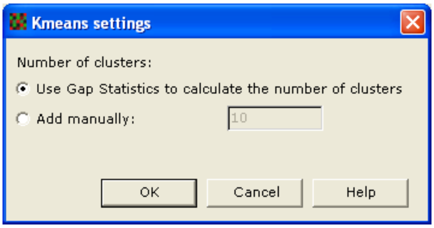
Calculation name: calculation 4

Fig E-9. EDA screenshot of K-means calculation setup

K-means can be calculated for proteins, spot maps and experimental groups by selecting the corresponding button.

Parameter	Description
Number of Clusters	No parameter is necessary for a K-means clustering in EDA, since EDA uses Gap statistics to calculate an optimal number of clusters for the dataset. If the user has prior knowledge that the data should be clustered into a specific number of clusters however, the number can be entered and the algorithm will be significantly speeded up.

Table E-4. Settings and parameters for K-means clustering analysis.



Kmeans settings

Number of clusters:

● Use Gap Statistics to calculate the number of clusters

○ Add manually: 10

Fig E-10. EDA screenshot of K-means settings dialog

Note that the distance measure in the K-means calculation is always Euclidean.

E.5 Self Organizing Maps

E.5.1 Introduction

Self Organizing Maps (SOM) is a method similar to K-means, but with the addition of organizing the clusters in a two dimensional map, where neighboring clusters show similar expression profiles.

- The SOM algorithm is relatively fast but not as fast as K-means.
- The neighboring clusters often have similar expression patterns since they are close to each other in the neuron layer and are thus moved in a similar way. The cluster results in EDA are presented in the same lattice as the algorithm uses.

E.5.2 Detailed Description

The data to be clustered in SOM defines the input layer, whereas the neuron layer contains the clusters which have relations to each other and to the data.

The aim of the learning steps (iterations) in SOM is to adapt the neuron layer to the input layer, in a similar fashion as the cluster centroids in K-means adapt to the data points.

There are, however, a few differences. This is the conceptual layout of the SOM algorithm:

1. Initially all the neurons are placed randomly in the input space with a reference vector.
2. The learning phase then begins. A random object from the input layer is chosen, say r. The distances from each node to this object and the distance between object and reference vector is then calculated using the selected similarity (distance) measure.
3. From these distances, the neuron with the closest distance is selected and called "best matching unit". This neuron's reference vector is then modified to move closer to the object
4. Then all neighboring neurons to the best matching unit are moved closer to the object (but by a smaller amount than the winning node). The farther away (topologically speaking) the nodes are from the winning node the less their reference vectors should be moved towards the input vector. The method that decides on the distance to move is called the neighborhood function.
5. Steps 2-4 are repeated until a certain number of iterations have been reached.

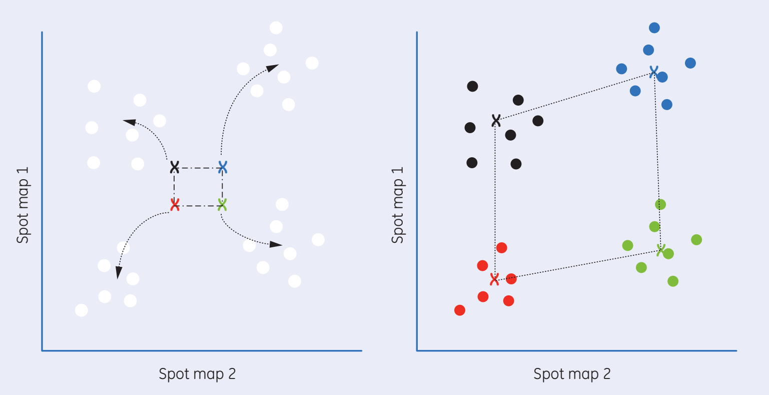


Fig E-11. The image describes the learning process in SOM.

Learning process and the neighborhood function

The learning process is an adaptive process that makes the two-dimensional lattice into an elastic net that stretches out over the input objects.

Only those neurons with that are topologically close to the best matching unit will learn from the same object.

The learning process can be defined as

$$w_j(n+1) = w_j(n) + \eta(n) h_{j,i(x)}(n)(x - w_j(n))$$

where **n** denotes the iteration, wj the reference vector of the **j**th neuron, **x** is the randomly selected object and **η(n)** is the learning rate parameter that will decrease with the number of iterations.

$\mathbf{h}_{\mathbf{j,i(x)}}$ is the neighborhood function that decides on how much each neuron **j** is to be moved relative the best matching unit **i**.

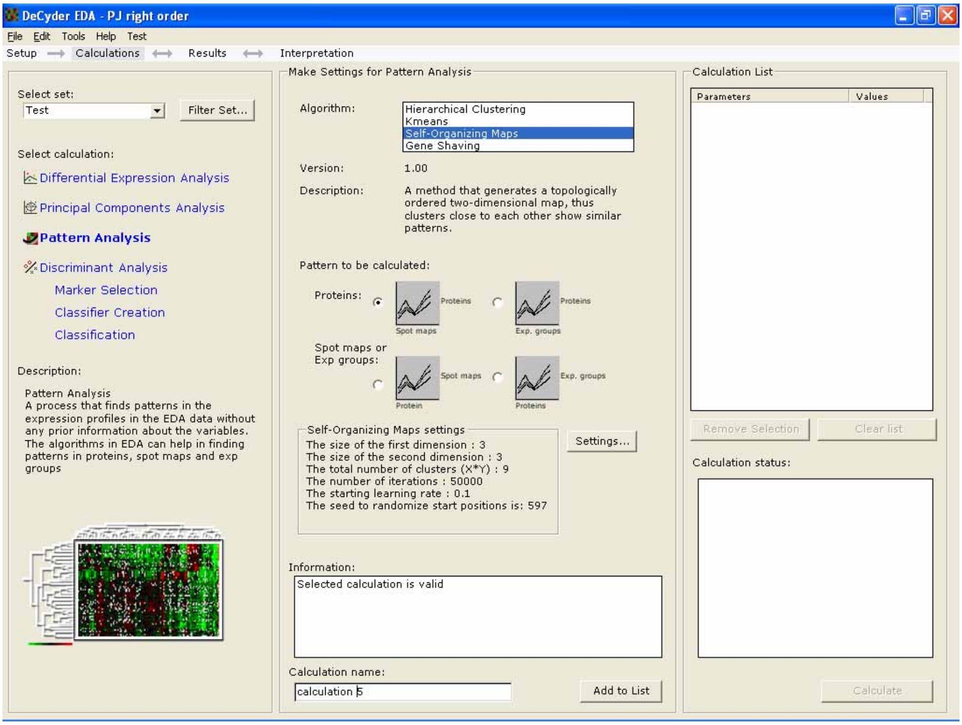
The neighborhood function, and the learning rate, decides how much each neuron will be moved in the direction of the object (relative to the best matching unit). The neighborhood function is designed to have a large value during the first iterations to rearrange nearly the whole lattice in every iteration, but after each iteration decreases in amount in order to let the lattice fine-tune itself.

Several different neighborhoods exist but in DeCyder EDA the function is a Gaussian function that can be described as:

$$h_{j,i(x)}(n) = \exp\left(\frac{d_{j,i}^2}{2\sigma^2(n)}\right)$$

where **σ(n)** is the width of the topological neighborhood that will decrease with the number of iterations.

E.5.3 Calculation Setup



Select set: Test

Select calculation: Pattern Analysis

Description:
Pattern Analysis
A process that finds patterns in the expression profiles in the EDA data without any prior information about the variables. The algorithms in EDA can help in finding patterns in proteins, spot maps and exp groups

Make Settings for Pattern Analysis

Algorithm: Hierarchical Clustering / Kmeans / Self-Organizing Maps / Gene Shaving

Version: 1.00

Description: A method that generates a topologically ordered two-dimensional map, thus clusters close to each other show similar patterns.

Pattern to be calculated:

Proteins: ● ○

Spot maps or Exp groups: ○ ○

Self-Organizing Maps settings
The size of the first dimension : 3
The size of the second dimension : 3
The total number of clusters (X*Y) : 9
The number of iterations : 50000
The starting learning rate : 0.1
The seed to randomize start positions is: 597

Information: Selected calculation is valid

Calculation name: calculation 5

Calculation List

Calculation status:

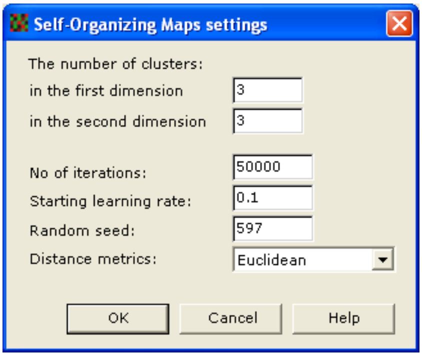
Fig E-12. EDA screenshot of Self-Organizing Maps calculation setup

Self-Organizing Maps can be calculated for proteins, spot maps and experimental groups by selecting the corresponding button.

Parameter	Description
The number of clusters	The number of clusters are defined as the number of neurons in each dimension in the two dimensional lattice.

Parameter	Description
The number of iterations	The number of iterations to use. The default value is 50000. A rule of thumb is to use at least 500 * the number of clusters.
The starting learning rate	The learning rate can also be adjusted, to optimize the learning process for different data. The default value is 0.1
The random seed	The random seed number initiates the random generator that defines where to put the initial neurons. If a test is to be reproduced with the same settings, the random seed must be the same.
Distance metrics	Choose Euclidean or Pearson Correlation. Euclidean is default.

Table E-5. Settings and parameters for Self-Organizing Maps clustering analysis.



Self-Organizing Maps settings

The number of clusters:

in the first dimension: 3

in the second dimension: 3

No of iterations: 50000

Starting learning rate: 0.1

Random seed: 597

Distance metrics: Euclidean

Fig E-13. EDA screenshot of Self-Organizing Maps settings dialog.

E.6 Gene Shaving

E.6.1 Introduction

Gene Shaving is a relatively new algorithm that was designed especially for expression analysis. The purpose of the algorithm is to identify groups of objects that have similar expression profiles and have optimal variation properties, meaning high variance between clusters but high coherence within the cluster.

Gene Shaving is not like other unsupervised algorithms due to the fact that the objects can be assigned to several clusters, making the clusters overlap, and that the sign of the expression value is disregarded, which may result in clusters with both linear object increase and decrease.

The name Shaving comes from the algorithm layout where a percentage of the objects are removed ("shaved off") during the different iterations.

- The Gene Shaving algorithm is relatively fast for clustering a large number of objects, but the calculation time increases rapidly with the number of observables and is usually regarded as slower than the K-means algorithm.
- The Gene Shaving algorithm finds overlapping clusters that are independent of each other.
- Objects that are clustered can be assigned to several clusters.
- The sign of the expression values to be clustered are disregarded which results in clusters containing both increasing and decreasing profiles. Only the absolute value is used.
- Gene Shaving may reveal structures other clustering algorithms cannot find.

E.6.2 Detailed Description

The Gene Shaving algorithm conceptually works like this:

The layout is described as the case where proteins are to be clustered.

1 The data expression matrix **X**, where the rows are proteins (variables) and the columns are our spot maps (observables), is row centered

2 The first principal component of the rows in matrix **X** is calculated.

3 A portion (α percent) of the proteins having the smallest loadings for the leading principal component is shaved off.

4 2) and 3) are repeated until no more objects are shaved off.

5 The result is a nested sequence of protein cluster results,

$$S_N \supset \ldots \supset S_k \supset \ldots \supset S_1$$

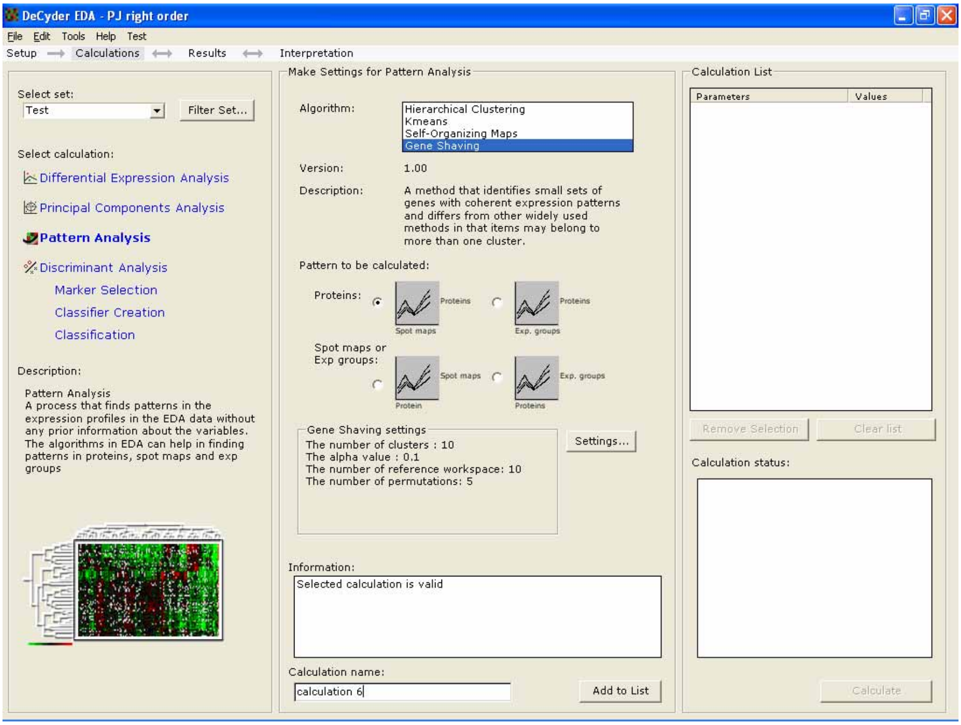
where the gene cluster $\mathbf{S_k}$ consists of **k** proteins.

The optimal cluster size is estimated using Gap Statistics (see below).

6 Each row of **X** is orthogonalized with respect to the average of the **k** proteins in the cluster with optimal size.

7 Steps 1-6 above are repeated with the orthogonalized data to find the next optimal cluster. This is repeated until a maximum of **M** clusters are found, with **M** chosen by the user.

E.6.3 Calculation Setup



Select set: Test

Select calculation: Pattern Analysis

Description:

Pattern Analysis
A process that finds patterns in the expression profiles in the EDA data without any prior information about the variables. The algorithms in EDA can help in finding patterns in proteins, spot maps and exp groups

Make Settings for Pattern Analysis

Algorithm: Gene Shaving

Version: 1.00

Description: A method that identifies small sets of genes with coherent expression patterns and differs from other widely used methods in that items may belong to more than one cluster.

Pattern to be calculated:

Proteins:

Spot maps or Exp groups:

Gene Shaving settings
The number of clusters : 10
The alpha value : 0.1
The number of reference workspace: 10
The number of permutations: 5

Information:
Selected calculation is valid

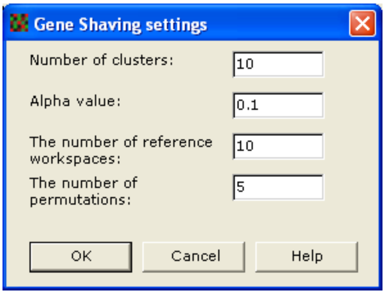
Calculation name: calculation 6

Fig E-14. EDA screenshot of Gene Shaving calculation setup

Gene Shaving can be calculated for proteins, spot maps and experimental groups by selecting the corresponding button.

Parameter	Description
Number of Clusters	Whether to use Gap Statistics to estimate the number of clusters or if the number should be entered manually.
Percentage to shave off (alpha value)	The alpha value indicates the percentage of objects that will be shaved off in each iteration, around 10 - 15 % is ideal.
The number of reference workspaces for the Gap Statistics calculation.	The larger the number the more time it will take to calculate, but more reference workspaces will constitute a better reference population. A value of 10 is used as default.
The number of permutations in the Gap Statistic calculation	This is the number of permutations that will be done to create a reference workspace. 5 is default.

Table E-6. Settings and parameters for Gene Shaving clustering analysis.



Gene Shaving settings

Number of clusters: 10

Alpha value: 0.1

The number of reference workspaces: 10

The number of permutations: 5

Fig E-15. EDA screenshot of Gene Shaving settings dialog.

E.7 Validation

E.7.1 Introduction

An important aspect of pattern analysis is the validation of the clusters and their quality.

The validity measures can be used to analyze which clustering algorithm results in the highest quality measure and how many clusters to divide the data into.

Method	Function	Application
Dunn's index	A measure of quality of the clusters in a result. The Dunn measure is used to compare different partition clustering results.	Is used to compare different clustering algorithms that have clustered the same data. For instance a SOM and K-means clustering.
Gap Statistics	Calculate an optimal number of clusters.	Is used in K-means to let the algorithm calculate the best number of clusters to use. Is also internally used in the Gene Shaving algorithm.

Table E-7. Validation methods for partitioned clustering analysis.

E.7.2 Dunn's Index

Dunn's Index is a quality measure that Dunn proposed in 1974. The measure attempts to find compact and well separated clusters by evaluating the distances within a cluster and between clusters.

$$D = \min_{1 \le i \le k} \left\{ \min_{\substack{i \le i \le k \\ i \ne j}} \left\{ \frac{d(C_i, C_j)}{\max\limits_{1 \le i \le k} \{\Delta(C_k)\}} \right\} \right\}$$

where **C**$_i$ is cluster **i**, d(**C**$_i$,**C**$_j$) is the inter cluster distance (between clusters) and Δ(**C**$_k$) is the intra cluster distance (within cluster).

It can easily be seen that well separated homogenous clusters have high inter-cluster distance and low intra-cluster distance. The conclusion is that large values of Dunn indicate compact and well-separated clusters.

The Dunn's index can be seen in EDA for the K-means, SOM and gene shaving algorithms as a setting in the calculation settings dialog.

E.7.3 Gap statistics in Gene Shaving and cluster validation

Gap Statistics is a quality measure for a cluster. Like the Dunn's index, it favors both high-variance clusters and high coherence between members of the cluster.

Gap Statistics in Gene Shaving

The algorithm uses variability in analogy with ANOVA which can be observed by the following measures for a cluster Sk. The assumption here is clustering of proteins.

Within variance

$$V_W = \frac{1}{p}\sum_{j=1}^{p}\left[\frac{1}{k}\sum_{i\in S_k}\left(x_{ij}-\bar{x}_j\right)^2\right]$$

Between variance

$$V_B = \frac{1}{p}\left(\bar{x}_j-\bar{x}\right)^2$$

Total variance

$$V_T = \frac{1}{kp}\sum_{i\in S_k}\sum_{j=1}^{p}\left(x_{ij}-\bar{x}\right)^2 = V_B + V_W$$

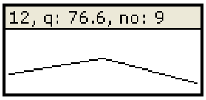
where the expression matrix **X** consists of **p** samples and **N** proteins with proteins on the rows and samples in the columns. $\mathbf{x_{ij}}$ is the log standard abundance of protein **i** in sample **j**.

The within variance measures the variability of each protein about the cluster average, averaged also over the samples. The VW can be small if the overall variance is small, so a more applicable measure is the VB/VW, the between-to-within variance ratio or if the value is transformed to the explained percent variance:

$$R^2 = 100\frac{V_B}{V_T} = \frac{\frac{V_B}{V_W}}{1+\frac{V_B}{V_W}}$$

A large **R^2** value implies a tight cluster of proteins.

DeCyder EDA uses the **R^2** value as a quality measure of a cluster.



12, q: 76.6, no: 9

Fig E-16. The quality value q (**R^2** value) for a cluster in EDA.

In the Gene shaving case, the optimal number of proteins in **S_k** is desired. The idea here is to see if the **R^2** value is larger than expected by chance. Therefore, **B** number of permuted data sets are created **X^{*b}**, by permuting the data in the rows of the expression matrix.

If **D_o** is the **R^2** value from the cluster and **$\overline{D}_k^{\ *}$** is the average **R^2** value for the permuted matrices, the gap function is defined as:

$Gap(k) = D_o - \overline{D}_k^{\ *}$

The optimal number of proteins is the value **k** that produces the largest gap.

Gap Statistics in estimating the number of clusters

The pair-wise distance between all observables in a cluster **Cr** can be measured with the squared Euclidean distance:

$$D_r = \sum_{i \in Cr} \sum_{j \in Cr} \left(x_{ij} - x_{i'j} \right)^2$$

Then the pooled within-cluster sum of squares around the cluster sum can be defined as:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

The idea is to test the log $\mathbf{W_k}$ against a null reference.

By introducing **B** reference data sets the following Gap Measure can be calculated:

$$Gap(k) = (1/B) \sum_{b} \log(W_{kb}^{*}) - \log(W_k)$$

The standard deviation $\mathbf{sd_k}$ is then defined as:

$$sd_k = \left[(1/B) \sum_{b} (\log(W_{kb}^{*}) - (1/B) \sum_{b} \log(W_{kb}^{*}))^2 \right]^{1/2}$$

and

$$s_k = sd_k \sqrt{1 + 1/B}$$

The optimal number of clusters is the smallest **k** such that *Gap(k) ≥ Gap(k+1) - sk+1*.

E.8 References

Cluster analysis

Everitt, B. S., Landau, S. and Leese, M. 2001. Cluster Analysis, 4th edition. Edward Arnold.

Eisen, MB; Spellman, PT; Brown, PO; Botstein, D (1998) "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci USA 95 14863-14868

Hierarchical Clustering

Sokal, R. & Mitchener, C. (1958), `A statistical method for evaluating systematic relationships', Univ. Kansas Sci. Bull.. 38, 1409--1438.

(for expression data, see Eisen 1998 above)

K-means

Lloyd, S. (1957), Least squares quantization in pcm., Technical report, Bell Laboratories. Published in 1982 in IEEE Trans. Inf. Theory, 28, 128137.

SOM

Kohonen (1990) The Self-Organizing Map Proc IEEE 78(9):1464-1480

P.Tamayo et al., "Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation," Proceedings of the National Academy of Sciences (USA) 96, No. 6, 2907-2912 (1999

Gene Shaving

Trevor Hastie, Robert Tibshirani, Michael Eisen, Patrick Brown, Doug Ross, Uwe Scherf, John Weinstein, Ash Alizadeh, Louis Staudt and David Botstein (2000). Gene Shaving: a New Class of Clustering Methods for Expression Arrays, Technical Report, Department of Statistics, Stanford University

Dunn

J. Dunn, "Well separated clusters and optimal fuzzy partitions", J.Cybernetics, Vol. 4, 1974, pp. 95-104

N. Bolshakova, F. Azuaje, Cluster validation techniques for genome expression data, Signal Processing, v.83 n.4, p.825-833, 2003

Gap Statistics

Tibshirani, R; Walther, G; Hastie., T "Estimating the Number of Clusters in a Dataset via the Gap Statistic." Technical report, Department of Biostatistics, Stanford University,

Appendix F Statistics and algorithms - Discriminant Analysis

F.1 Introduction

In real life classification or discriminant analysis happens on a regular basis, for instance when selecting which fruit to buy in the supermarket.

The fruit is classified depending on certain properties or features, e.g. on the ripeness of the fruit, the color of the fruit or how soft the fruit is. So, by investigating different features, the fruit can be classified as good or bad.

In data mining, when one wants to see which property or feature can discriminate between different classes, for instance good or bad as in the fruit case, a process called Feature Selection is used. Feature Selection, or as it is called in EDA, Marker Selection, is a selection process to see if the variable is useful or not.

During our lifetime, sufficient experience has been gained to say that different levels of the fruit features correspond to if the fruit is good or bad. We have, by learning, actually built a model of the reality, or a classifier that predicts the class or category, based on certain levels.

This process is called Model building or Classifier Creation.

So to decide if it is good or bad fruit, another method called Classification is used. The item is categorized into a predefined class. In the fruit example the classes are "good" and "bad".

F.2 In EDA

In DeCyder EDA, these three mechanisms exist to help facilitate different aspects of the analysis. However only the spot maps can be classified, and thus be used as observables. The class labels of these spot maps can either be experimental groups or conditions and contain information about e.g. different tumor types, rate of disease progression or response to therapy, to name but a few.

Since an EDA workspace can consist of thousands of proteins and since the number of proteins (features) almost always is greater then the number of samples, it is likely that most proteins will not carry relevant information for discrimination between the classes. A goal is therefore to find as small subset as possible of the proteins that can discriminate between the classes. In addition, a small set of proteins is desirable for diagnostics and prognostic purposes.

It might be of interest to predict or classify unknown samples to a predefined class, for prognostic or diagnostic purposes. By creating models and by analyzing the accuracy of each model, the best model can be selected and used for classification of an unknown sample.

F.3 Training and Testing

When creating a classifier in EDA, EDA measures a classifier's performance in prediction accuracy (see below). EDA cannot use the training set for testing the performance of the classifier, since the model is biased versus that data. The training data was used during the learning process to determine the parameters of the classifier, thus the model has already seen that data. Given that the classifier probably was built for use on other data, the real performance must be tested on data it hasn't seen before, the independent test data.

An important assumption here is that both the training and the test set are representative for the whole problem domain.

For example, if a scientist builds a hypothetical classifier for detecting if wine comes from a certain wine district or not, the scientist wouldn't just use wine from a single wine distributor in the district as a positive training set, since that wouldn't represent the whole wine district. So when creating a classifier the training set needs to be as generalized as possible to include the different variations that may exist within the different classes.

F.3.1 Training and Test sets

If the dataset is large enough, two independent sets of data can be used, one as training and one for calculating the performance accuracy, the test set. If the test set is a good representative, the accuracy is a good measurement for future performance.

A large training set will probably cover more of the problem domain and give better models, whereas a large test set will give a good performance estimation.

In biological applications, the number of samples is often limited and another approach is to use all samples since one cannot afford to hold back samples. But the performance must still be estimated on an independent test set.

F.3.2 Cross Validation

When the dataset isn't large enough and can't be divided into an independent test and training set since all data must be used, cross validation can be useful.

In Cross Validation (CV), a fixed number of folds are decided. Then the data is divided into that number of folds using approximately equal sizes. In EDA, the data is also divided using stratification so that when the number of folds are small enough, each class is represented in each fold. If stratification isn't used, the training and test set are not well represented.

In EDA, the stratification process is done automatically.

The CV process can be described as follows:

1 Divide the data randomly into k stratified folds.

2 Train the model on k-1 folds, use one fold for testing.

3 Repeat the process k times so that all folds are used for testing.

4 Compute the average performance on the k tests.

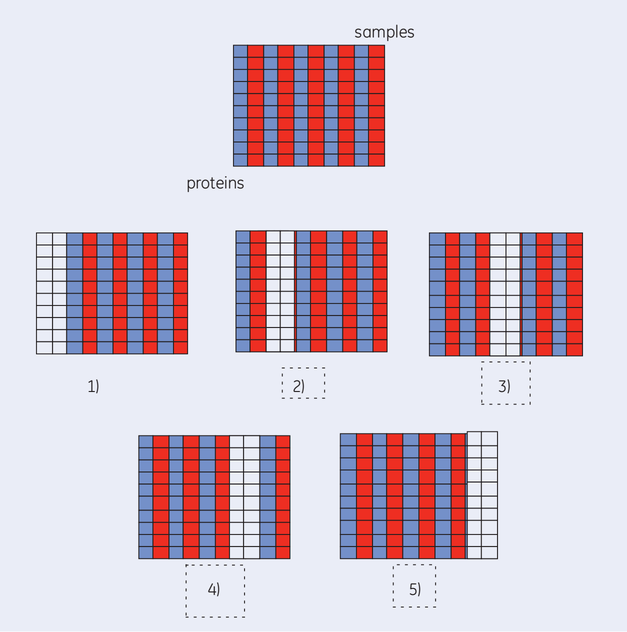


Fig F-1. A hypothetical dataset containing 10 proteins and 10 samples, where the spot maps belong to either the blue or red class. In a 5 fold classification different parts of the data set are used for training and testing (grey).

Since every repeat generates a classifier, CV generates as many independent classifiers as the number of folds.

Leave-one-out cross validation is commonly used, and the principle there is to have as many CV folds as samples, which means that during each repeat a single sample is left out of the training session and later on used as a test set. In EDA, Leave-one-out CV can be accomplished by setting the number of folds to the maximum number.

F.3.3 Prediction accuracy

An ordinary prediction accuracy is calculated as:

Accuracy = Number of correct predictions / Number of predictions

For a two-class problem of positive and negative samples:

	Positive	**Negative**
Predicted as positives	TP	FP
Predicted as negative	FN	TN

the accuracy will be calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

but since the classes can be unbalanced, EDA has corrected for that by using:

Accuracy = average of (correct predictions / number of predictions) for each class

Example

If a classifier was used to classify 5 spot maps that have known classes:

True Class	**Predicted Class**
Class 1	Class 2
Class 2	Class 1
Class 3	Class 3
Class 3	Class 3
Class 3	Class 3

With traditional accuracy the classifier would get an accuracy of 60%, even if it missed two classes completely. With the weighted accuracy in EDA the result will be 1/3 (0/1 + 0/1 + 3/3) = 33%, which is a better measure of this classifier.

F.4 Feature / Marker Selection

F.4.1 Introduction

Several search methods exist to select a subset of features from the whole training set. In EDA this corresponds to selecting the proteins from a set that can best discriminate the spot maps in the test set.

The selection problem is a complex problem since the number of possible feature subsets increases exponentially with the number of features.

The number of possible subsets is 2^p where p is the number of features.

So if there are 3 proteins (P1, P2, and P3), $2^3 = 8$ feature subsets can be generated.

(P1, P2, P3, P1P2, P1P3, P2P3, P1P2P3, no P).

In the extreme case (Exhaustive search) all combinations are tested to see which combination gives the highest performance. Due to the time it takes to do these calculations two of the few other approaches that exist are also implemented in DeCyder EDA:

Method	Function	Application
Forward Selection	Divides the proteins in the set into different subsets in an iterative way by adding protein after protein.	Is used to find the best subset of proteins that discriminates between the different classes.
Partial Least Squares Search	Calculates a Partial Least Squares and from the result tries to find the best proteins to use.	Is used to find the best subset of proteins that discriminates between the different classes.

Table F-1. Overview of methods used in Feature / Marker Selection.

In EDA a feature is a protein and depending on the expression levels in the classes it can be a good or bad feature for discriminating between the classes.

Example

In an experiment there is expression data for three proteins from samples belonging to either a normal or a tumor tissue. Are any of the proteins a good feature from a discriminative point of view?

	Normal 1	Normal 2	Tumor 1	Tumor 2
Protein 1	0.06	0.04	0.12	0.07
Protein 2	0.03	0.03	-0.97	-0.96
Protein 3	-0.03	-0.03	1.12	1.13

Table F-2. Protein expression data for feature selection.

Protein 2 is down-regulated in the tumor samples compared to the normal and one can clearly distinguish the two classes from this protein. Protein 3 also shows differences in the expression between the two groups but is up-regulated in the tumor type compared to the normal.

Often a single feature is not sufficient. Instead, a set of features is necessary for a good classification. For example, by classifying people by gender, the weight or the length of the people can give some information but these two variables cannot be used for a 100% correct classification. More information, thus more variables are needed.

The same is true about protein expression in, for example, a multi-class problem. Some proteins might discriminate between some of the classes, some might discriminate between some other classes, and altogether they might discriminate between all the classes, depending on their values for each respective class.

F.4.2 Detailed Description

Strategies for Feature Selection

There exist two types of strategies for feature selection.

1. Filter approach. Features are selected independent of the learning algorithm.

2. Wrapper approach. Measures performance with the learning algorithm to select and evaluate feature set.

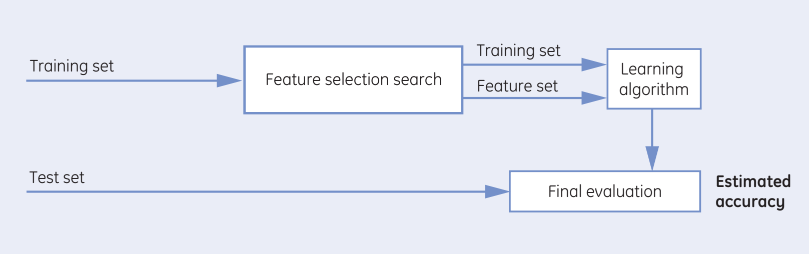


Fig F-2. A schematic filter approach where the features are selected independently of the learning algorithm, for instance by a univariate method such as One-Way ANOVA.

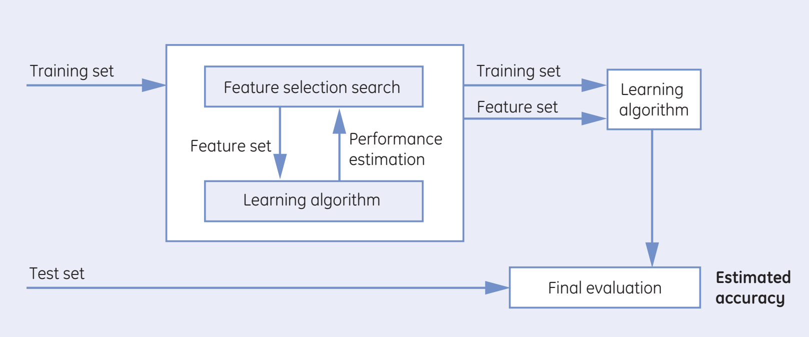


Fig F-3. A schematic wrapper approach where the learning algorithm has been wrapped inside the selection method.

Forward Selection

Forward selection is a wrapper method, which iteratively selects a new feature and thereby creates a feature set that is most likely to best predict the class.

The process:

1 Start with no features and create p subsets containing one feature in each and then send them to the learning algorithm to see which of all the features has the highest accuracy when using just that feature. Keep the highest scoring feature in the total feature set.

2 Create new feature sets by taking the total feature set and adding one of the rest.

3 Send the new feature sets to the learning algorithm to get the accuracy. Keep the highest scoring feature set as the total feature set.

4 Send the new feature sets to the learning algorithm to get the accuracy. Keep the highest scoring feature set as the total feature set.

The forward selection creates a locally optimal feature set but not necessarily globally, since it doesn't try all possibilities as the exhaustive search does.

The result of the Forward Selection is an accuracy graph and two lists of the rank and the appearance of the proteins.

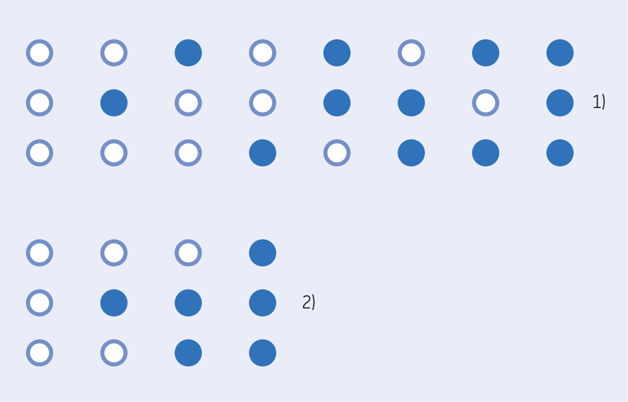
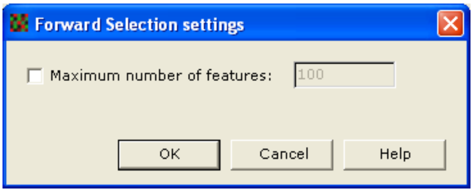


Fig F-4. An example of 3 proteins that are selected (black) or not selected (white) in Exhaustive search (1) and Forward selection (2). For Exhaustive search all 8 possible states are check for the best combination. Forward selection finds the best protein for each step.


Forward Selection settings

□ Maximum number of features: 100

Fig F-5. EDA screenshot of Forward Selection settings dialog.

Parameter	Description
Number of Features	Select the maximum number of features to test. Sometimes it can be a good idea to set a maximum number of features to search if one is only interested in a few numbers.

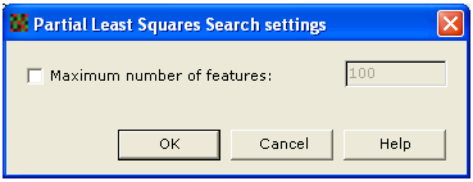
Table F-3. Settings and parameters for Forward Selection calculation.

Partial Least Squares Search

Partial Least Squares (PLS) Search is a novel algorithm that creates a PLS model of the data and then uses the Variable Influence on the Projection (VIP) scores from the model to create a ranked list of how good the features are for discrimination between the classes.

The features are then added to the total feature set from the top of the list. As with Forward selection, the result of the PLS Search is an accuracy graph and two lists of the rank and the appearance of the proteins.

For more information about PLS and VIP see reference list.


Partial Least Squares Search settings

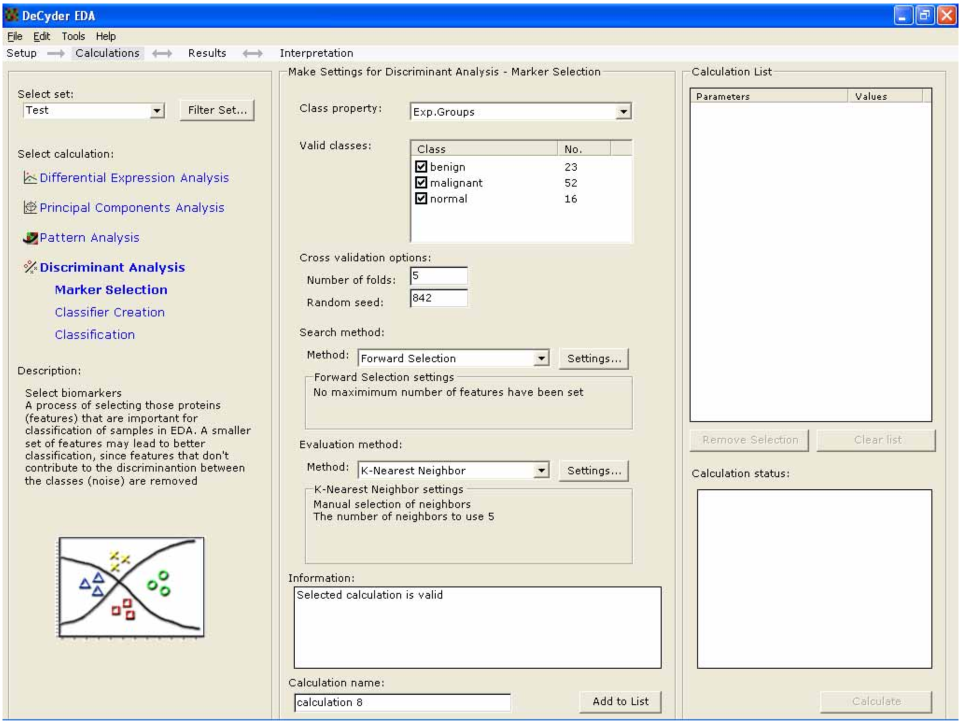
□ Maximum number of features: 100

Fig F-6. EDA screenshot of Partial Least Squares Search settings dialog.

Parameter	Description
Number of features	Select the maximum number of features to test. Sometimes it can be a good idea to set a maximum number of features to search if one is only interested in a few numbers.

Table F-4. Settings and parameters for Partial Least Squares Search calculation.

F.4.3 Calculation Setup



Select set: Test

Select calculation:
Differential Expression Analysis
Principal Components Analysis
Pattern Analysis
Discriminant Analysis
Marker Selection
Classifier Creation
Classification

Description:
Select biomarkers
A process of selecting those proteins (features) that are important for classification of samples in EDA. A smaller set of features may lead to better classification, since features that don't contribute to the discriminantion between the classes (noise) are removed

Make Settings for Discriminant Analysis - Marker Selection

Class property: Exp.Groups

Valid classes:

Class	No.
☑ benign	23
☑ malignant	52
☑ normal	16

Cross validation options:
Number of folds: 5
Random seed: 842

Search method:
Method: Forward Selection
Forward Selection settings
No maximimum number of features have been set

Evaluation method:
Method: K-Nearest Neighbor
K-Nearest Neighbor settings
Manual selection of neighbors
The number of neighbors to use 5

Information:
Selected calculation is valid

Calculation name: calculation 8

Calculation List

Parameters	Values

Calculation status:

Fig F-7. EDA screenshot of Marker Selection calculation setup.

Parameter	Description
Class definitions	Select the parameter that decides the classes. The different classes can be • Experimental Groups • One of the conditions defined. If a class shouldn't be used but is present in the set, uncheck the box for that class. The spot maps belonging to that class will not be used in the feature selection.
Number of Folds	Select the number of folds for the cross validation
Search Method	Select the search method for marker selection
Evaluation (Classification) Method	Select the learning method (classifier) for marker selection

Table F-5. Settings and parameters for Marker Selection calculation.

F.5 Classifier Creation

F.5.1 Introduction

There are several learning algorithms or supervised learning methods that can be used to build models.

Method	Function	Application
K-Nearest Neighbors (KNN)	Calculates the distance between an unknown spot map and the training data and classifies the unknown spot map to the majority class label of the k nearest neighbors in space.	Is used to build a classifier and for accuracy estimations in Marker selection.
Regularized Discriminant Analysis (RDA)	Calculates the posterior probability of the unknown spot map to the classes in the training data and classifies the unknown spot map to the class with the highest probability.	Is used to build a classifier and for accuracy estimations in Marker selection.

Table F-6. Overview of methods used in Classifier Creation.

F.5.2 Detailed Description

K-Nearest Neighbors

The K-Nearest Neighbor (KNN) is a simple classifier that doesn't create any rules or weights when introduced to the learning data. Instead, the classifier stores the training data and the calculation is performed when it comes to classifying new data (test data or unknown data).

In the KNN case each new sample that is to be classified is compared to the training data using a distance measure and the unknown sample are classified into the same class as the closest sample in the training set belongs to.

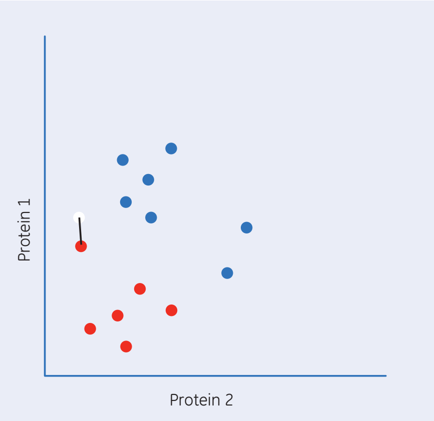


Fig F-8. In a 1-KNN the closest training data according to the similarity measure is used to predict the class the unknown sample (grey) belongs to. In this case the sample is assigned to the red class.

Often more than one sample is used to assign the class of the unknown sample. In those cases the majority class of the samples closest to the unknown sample is selected.

The k in KNN is the number of samples to use for the class assignment.

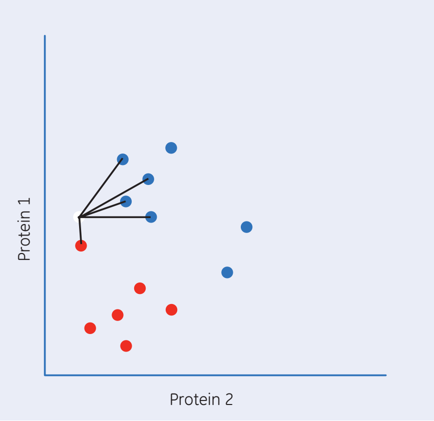
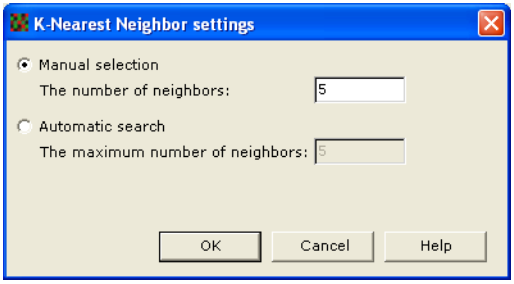


Fig F-9. In a 5-KNN the 5 closest training data according to the similarity measure are used to predict the class the unknown sample (grey) belongs to. In this case the sample is predicted to be a member of the blue class, by majority vote.



K-Nearest Neighbor settings

● Manual selection

The number of neighbors: 5

○ Automatic search

The maximum number of neighbors: 5

Fig F-10. EDA screenshot of K-Nearest Neighbor settings dialog.

Parameter	Description
Manual or Automatic selection of settings	In manual selection the classifier uses the specified setting. In automatic selection, the classifier tries several setting possibilities and stores the best result. The drawback with automatic selection is that is time-consuming.
The number of neighbors (k-value)	It can thus be seen that the k-value should be chosen carefully and not set to a higher number than the smallest class in the training data. By default, DeCyder EDA automatically suggests the highest number possible.
The maximum number of neighbors (k-value)	When doing an automatic selection the algorithm needs to have a maximum number of neighbors to test. The algorithm will start at 1 and test all the way up to this number.

Table F-7. Settings and parameters for K-Nearest Neighbor calculation.

Regularized Discriminant Analysis

Traditional Discriminant Analysis (DA) methods are often used for classification problems and a general assumption for these supervised algorithms is that they assume Gaussian distribution of the classes. The classifier is thus based on the Gaussian (normal) distribution.

$$p(x \mid \omega_i) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) \right\}$$

where **$p(x \mid \omega_i)$** is the class density function, where ω_i is class **i** and **m_i** is the mean of the class **i**.

Σ_i is the covariance matrix of class **i**.

The original DA method, Linear Discriminant Analysis (LDA), was introduced by Fischer and assumes that the classes have different class means but identical covariance, which leads to linear decision borders between the classes.

If one instead assumes that the covariance matrices are different one gets Quadratic Discriminant Analysis (QDA) instead. The different covariance matrices lead to non-linear decision boundaries.

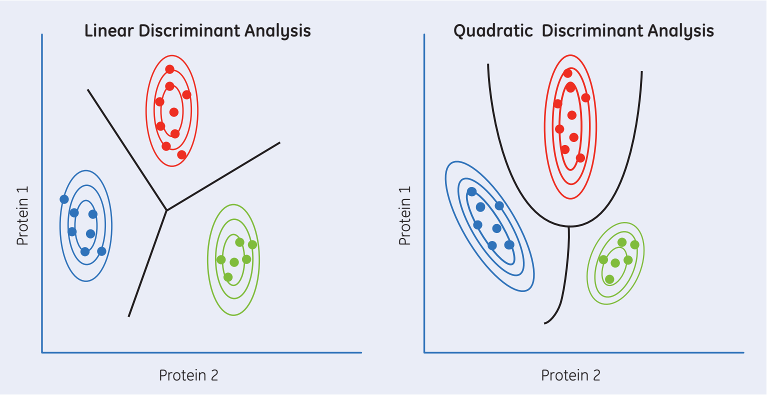


Fig F-11. The difference in the assumption of covariance matrix enables the QDA to have non-linear decision boundaries.

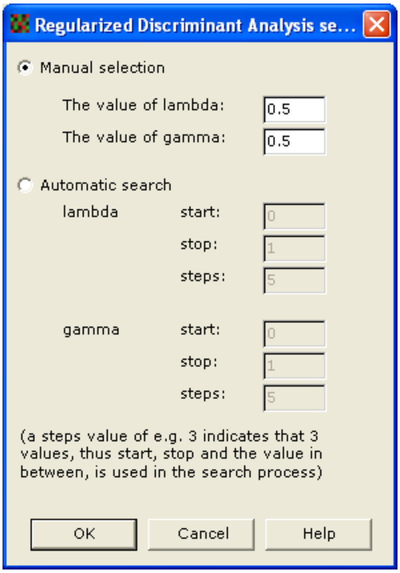
Since the decision boundaries in LDA are linear it is sometimes not flexible enough whereas the QDA is less stable in some cases. Therefore, Regularized Discriminant Analysis was introduced by Friedman in 1989 as a compromise between LDA and QDA to overcome their drawbacks.

Using a parameter alpha the covariance matrix can be shifted towards LDA or QDA.

lambda = 0 gives QDA.

lambda = 1 gives LDA.

The second parameter gamma is used to regularize the sample covariance matrix to overcome the quadratic stability problem.



Regularized Discriminant Analysis se...

● Manual selection

The value of lambda: 0.5

The value of gamma: 0.5

○ Automatic search

lambda start: 0

stop: 1

steps: 5

gamma start: 0

stop: 1

steps: 5

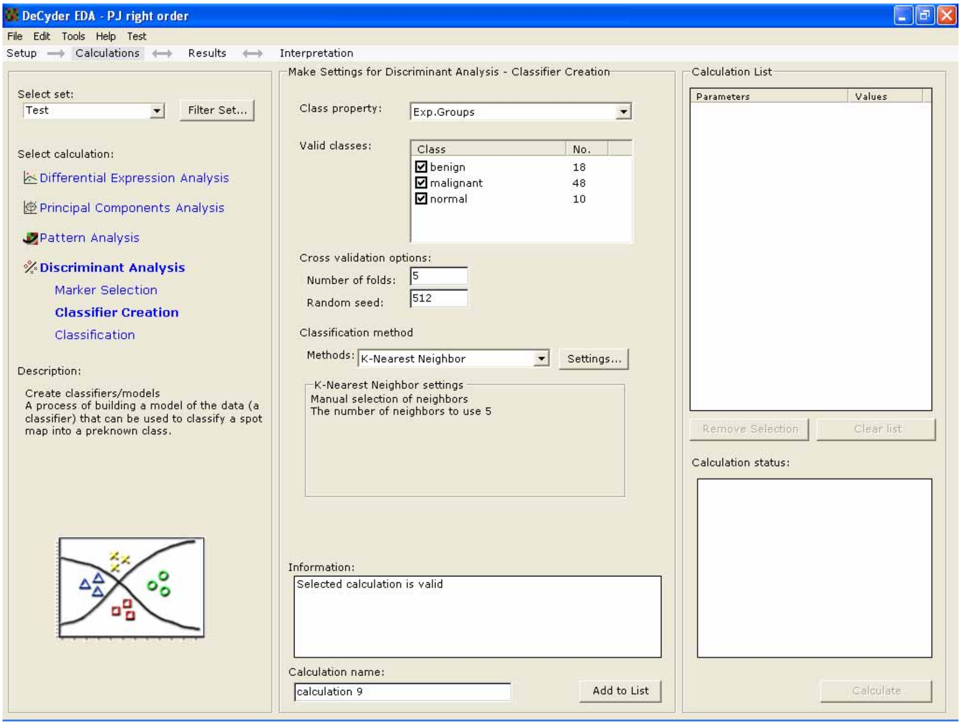
(a steps value of e.g. 3 indicates that 3 values, thus start, stop and the value in between, is used in the search process)

Fig F-12. EDA screenshot of Regularized Discriminant Analysis settings dialog.

Parameter	Description
Manual or Automatic selection of settings	In manual selection the classifier uses the specified setting. In automatic selection, the classifier tries several setting possibilities and stores the best result. The drawback with automatic selection is that it is time-consuming.
Lambda	A value closer to 0 gives more QDA-like decision borders. A value closer to 1 gives more LDA-like decision borders.
Gamma	Gamma is used to regularize the sample covariance matrix to overcome the quadratic stability problem.
Lambda start	The start value for lambda in automatic selection.
Lambda stop	The stop value for lambda in automatic selection.
Lambda steps	The number of steps to test in automatic selection. If start is 0, stop is 1 and number of steps is 6:0, 0.2, 0.4, 0.6, 0.8 and 1.0 are tested.
Gamma start	The start value for gamma in automatic selection.
Gamma stop	The stop value for gamma in automatic selection.
Gamma steps	The number of steps to test in automatic selection. If start is 0, stop is 1 and number of steps is 6:0, 0.2, 0.4, 0.6, 0.8 and 1.0 are tested.

Table F-8. Settings and parameters for Regularized Discriminant Analysis.

F.5.3 Calculation Setup


DeCyder EDA - PJ right order
File Edit Tools Help Test
Setup ⟶ Calculations ⟷ Results ⟷ Interpretation
Filter Set...
Settings...
Add to List
Remove Selection
Clear list
Calculate

Select set: Test

Select calculation:

- Differential Expression Analysis
- Principal Components Analysis
- Pattern Analysis
- **Discriminant Analysis**
  - Marker Selection
  - **Classifier Creation**
  - Classification

Description:

Create classifiers/models
A process of building a model of the data (a classifier) that can be used to classify a spot map into a preknown class.

Make Settings for Discriminant Analysis - Classifier Creation

Class property: Exp.Groups

Valid classes:

Class	No.
☑ benign	18
☑ malignant	48
☑ normal	10

Cross validation options:

Number of folds: 5

Random seed: 512

Classification method

Methods: K-Nearest Neighbor

K-Nearest Neighbor settings
Manual selection of neighbors
The number of neighbors to use 5

Information:
Selected calculation is valid

Calculation name: calculation 9

Calculation List

Parameters	Values

Calculation status:

Fig F-13. EDA screenshot of Model Creation calculation setup

Parameter	Description
Class definitions	Select the parameter that decides the classes. The different classes can be: • Experimental Groups • One of the conditions defined. If a class shouldn't be used but is present in the set, uncheck the box for that class. The spot maps belonging to that class will not be used in the feature selection.
Number of Folds	Select the number of folds for the cross validation
Classification Method	Select the classifying method

Table F-9. Settings and parameters for Classifier Creation.

F.6 Classification

Classification is achieved by assigning a sample to a class for which the posterior probability **p(x I ω_i)** is the greatest. Using Bayes' rule and the fact that **p(x)** is independent of class, the following decision rule applies:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\left[\Sigma_i^{\alpha,\gamma}\right]^{-1}(x-\mu_i) - \frac{1}{2}\log\left(\left|\Sigma_i^{\alpha,\gamma}\right|\right) + \log(p(\omega_i))$$

So to classify, the sample **x** is predicted to belong to the class **i** where **g$_i$(x)** is the highest.

F.7 References

Feature Selection and Classification

Statistical Pattern Recognition, Andrew Webb, 2nd edition 2002, Wiley

Data Mining, Witten I.H, Frank E., 2000, Morgan Kaufman.

Classification on expression data

Golub T et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286:531-537, 1999.

S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. June 2000.

Nguyen DV and Rocker DM. Tumor classification by partial least squares using microarray gene expression data. 2001.

Appendix G DiscoveryHub

The discoveryHub software is used for advanced biological and biochemical queries. The software searches various databases and collects data and organizes the search result in a predefined structure.

The **discoveryHub** software must be installed to be able to create PubMed queries in the **Interpretation** step of EDA. A license for the software must be purchased from GE Healthcare. Please contact GE Healthcare for more information.

When the software has been installed, settings for the connection to discovery Hub must be entered in the Database Administration Tool. See section G.2, Enter settings for discoveryHub for information.

Also, if clients and servers access the internet through a proxy server, proxy access must be enabled and proxy settings entered.

For detailed information on how to use discoveryHub, see instructions for discoveryHub software and instructions for applicable databases and servers.

G.1 Open Database Administration Tool

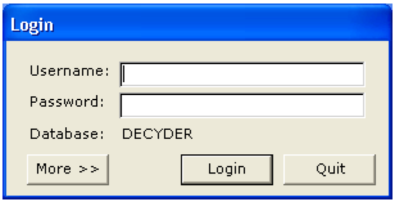
The DeCyder 2D database is administered from the DeCyder 2D Database Administration Tool and **Administrator** rights are required for performing **Database Administration tool** functions.

Note: *The DeCyder 2D Database Administration Tool can only be run on the computer where the DeCyder 2D database is installed.*

To open the Database Administration Tool:

1 In the **Start** menu of the computer with the DeCyder 2D database installed, select **All Programs:DeCyder 2D6 Software:Database Administration Tool**.

The **Login** dialog appears.



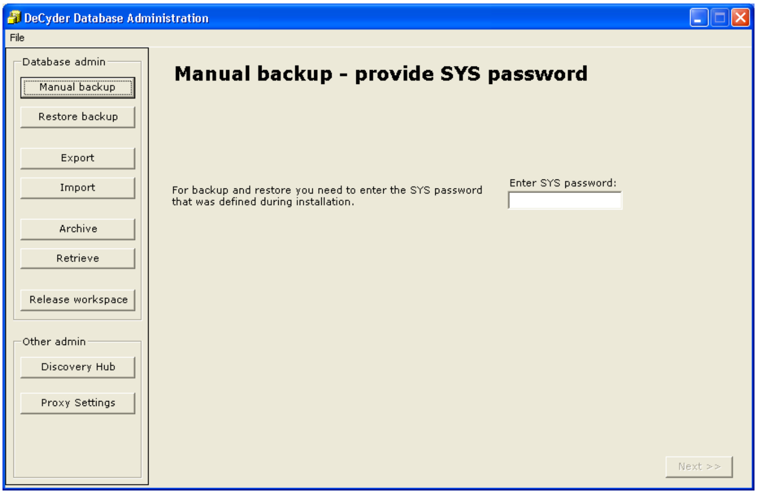
Database: DECYDER

2 Enter **User name** and **Password** and click **OK.** The DeCyder 2D database is set as default.

Tip! *To change database, click More >> and select another database.*

Note: *Only users in the group* ***User Administrator*** *can open and use the Database Administration Tool.*

3 The **DeCyder 2D Database Administration** main screen opens.

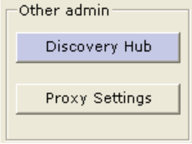


Manual backup - provide SYS password

For backup and restore you need to enter the SYS password that was defined during installation.

Enter SYS password:

G.2 Enter settings for discoveryHub

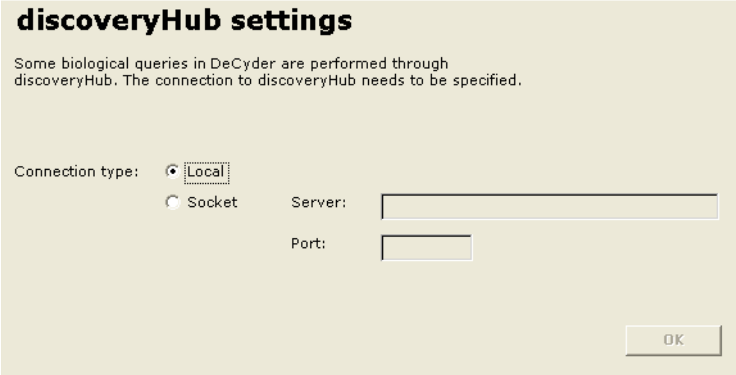


Other admin

Discovery Hub

Proxy Settings

1 Select **Discovery Hub** in the **Other admin** (other administration tools) in the **DeCyder 2D Database administration** main screen. The settings for discoveryHub are displayed.



discoveryHub settings

Some biological queries in DeCyder are performed through discoveryHub. The connection to discoveryHub needs to be specified.

Connection type: ● Local

○ Socket Server:

Port:

2 Select the type of connection that will be used to access discoveryHub (i.e. if discoveryHub is installed on your computer or on another computer that can be found on the network).

Examples

If you are not working on the computer where the database is installed but you have discoveryHub installed on your computer:

a. Go to the computer with the database and open Database Administration Tool on that computer.

b. Choose **Socket** in the discoveryHub settings area and enter your computer as server in the **Server** field. Enter the Port number.

If discoveryHub is installed on the same computer as the database but you work on another computer:

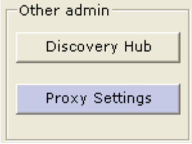
a. Go to the computer with the database and open Database Administration Tool on that computer.

b. Choose **Local** in the discoveryHub settings area.

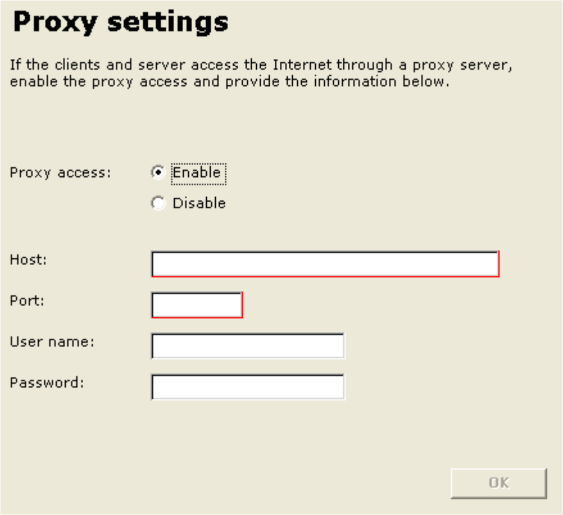
3 Click **OK** to save the settings.

G.3 Enable proxy access and enter proxy settings

If you use a proxy server to access the Internet (check with your system administrator), proxy access must be enabled and proxy settings entered.



1 Click **Proxy Settings** in the **Other admin** area. The **Proxy settings** area is displayed.



Proxy settings

If the clients and server access the Internet through a proxy server, enable the proxy access and provide the information below.

Proxy access: ● Enable ○ Disable

Host:

Port:

User name:

Password:

2 Select **Enable** to enable proxy access.

3 Enter the **Host** and **Port** (check with your system administrator if you do not know what to enter).

4 If required, enter **User name** and **Password** for the proxy access (check with your system administrator if user name and password are required).

Index

www.gehealthcare.com

GE Healthcare
Amersham Biosciences AB
Björkgatan 30
751 84 Uppsala
Sweden

All goods and services are sold subject to the terms and conditions of sale of the company within GE Healthcare which supplies them. GE Healthcare reserves the right, subject to any regulatory and contractual approval, if required, to make changes in specifications and features shown herein, or discontinue the product described at any time without notice or obligation. Contact your local GE Healthcare representative for the most current information.

CyDye: 2-D Fluorescence Difference Gel Electrophoresis (2-D DIGE) technology is covered by US patent numbers US6,043,025, US6,048,982, US6,127,134, and US6,426,190 and foreign equivalents and exclusively licensed from Carnegie Mellon University.

CyDye: this product or portions thereof is manufactured under licence from Carnegie Mellon University under US patent number US5,268,486 and other patents pending.

The purchase of CyDye fluors includes a limited license to use the CyDye fluors for internal research and development, but not for any commercial purposes. A license to use the CyDye fluors for commercial purposes is subject to a separate license agreement with GE Healthcare.

GE Healthcare has patent applications pending relating to its DeCyder software technology, European patent application number EP1,234,280.

Amersham Biosciences AB, a General Electric company going to market as GE Healthcare.

GE Healthcare Amersham Biosciences AB
Björkgatan 30, 751 84 Uppsala, Sweden

GE Healthcare Amersham Biosciences Europe GmbH
Munzinger Strasse 9, D-79111 Freiburg, Germany

GE Healthcare Amersham Biosciences UK Ltd
Amersham Place, Little Chalfont, Buckinghamshire, HP7 9NA, UK

GE Healthcare Amersham Biosciences Corp
800 Centennial Avenue, P.O. Box 1327, Piscataway, NJ 08855-1327, USA

GE Healthcare Amersham Biosciences KK
Sanken Bldg. 3-25-1, Hyakunincho, Shinjuku-ku, Tokyo 169-0073, Japan

Asia Pacific Tel: +852 2811 8693 Fax: +852 2811 5251 • Australasia Tel: + 61 2 9899 0999 Fax: +61 2 9899 7511 • Austria Tel: 01/57606-1619 Fax: 01/57606-1627 • Belgium Tel: 0800 73 888 Fax: 03 272 1637 • Canada Tel: 800 463 5800 Fax: 800 567 1008 • Central, East, & South East Europe Tel: +43 1 982 3826 Fax: +43 1 985 8327 • Denmark Tel: 45 16 2400 Fax: 45 16 2424 • Finland & Baltics Tel: +358-(0)9-512 39 40 Fax: +358 (0)9 512 39 439 • France Tel: 01 69 35 67 00 Fax: 01 69 41 96 77 • Germany Tel: 0761/4903-490 Fax: 0761/4903-405 • Italy Tel: 02 27322 1 Fax: 02 27302 212 • Japan Tel: +81 3 5331 9336 Fax: +81 3 5331 9370 • Latin America Tel: +55 11 3933 7300 Fax: +55 11 3933 7304 • Middle East & Africa Tel: +30 210 9600 687 Fax: +30 210 9600 693 • Netherlands Tel: 0165 580 410 Fax: 0165 580 401 • Norway Tel: 815 65 555 Fax: 815 65 666 • Portugal Tel: 21 417 7035 Fax: 21 417 3184 • Russia & other C.I.S. & N.I.S Tel: +7 (095) 232 0250, 956 1137 Fax: +7 (095) 230 6377 • South East Asia Tel: 60 3 8024 2080 Fax: 60 3 8024 2090 • Spain Tel: 93 594 49 50 Fax: 93 594 49 55 • Sweden Tel: 018 612 1900 Fax: 018 612 1910 • Switzerland Tel: 0848 8028 12 Fax: 0848 8028 13 • UK Tel: 0800 616928 Fax: 0800 616927 • USA Tel: 800 526 3593 Fax: 877 295 8102

imagination at work

ProTang Teknikinformation AB, Uppsala. Printed by Elanders Tofters 2005.